Lossy joint source-channel coding in the finite blocklength regime

Victoria Kostina, Sergio Verdú

Dept. of Electrical Engineering, Princeton University, NJ 08544, USA

Abstract

This paper shows new tight finite-blocklength bounds for the best achievable lossy joint source-channel code rate, and demonstrates that joint source-channel code design brings considerable performance advantage over a separate one in the non-asymptotic regime. A joint source-channel code maps a block of k source symbols onto a length$-n$ channel codeword, and the fidelity of reproduction at the receiver end is measured by the probability ϵ that the distortion exceeds a given threshold d. For memoryless sources and channels, it is demonstrated that the parameters of the best joint source-channel code must satisfy $nC - kR(d) \approx \sqrt{nV + k\mathcal{V}(d)}Q^{-1}(\epsilon)$, where C and V are the channel capacity and channel dispersion, respectively; $R(d)$ and $\mathcal{V}(d)$ are the source rate-distortion and rate-dispersion functions; and Q is the standard Gaussian complementary cdf. Symbol-by-symbol (uncoded) transmission is known to achieve the Shannon limit when the source and channel satisfy a certain probabilistic matching condition. In this paper we show that even when this condition is not satisfied, symbol-by-symbol transmission is, in some cases, the best known strategy in the non-asymptotic regime.

Index Terms

Achievability, converse, finite blocklength regime, joint source-channel coding, lossy source coding, memoryless sources, rate-distortion theory, Shannon theory.

This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1016625 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370. The first author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

I. Introduction

For a large class of sources and channels, in the limit of large blocklength, the maximum achievable joint source-channel coding (JSCC) rate compatible with vanishing excess distortion probability is characterized by the ratio $\frac{C}{R(d)}$ [1], attainable by separate source-channel coding (SSCC). However, at finite blocklengths not only is the fundamental limit $\frac{C}{R(d)}$ no longer achievable but separation ceases to be optimal. Quantifying, as a function of blocklength and excess distortion probability, the required backoff from $\frac{C}{R(d)}$ as well as the increase in the rate of source symbols per channel use afforded by an optimal joint design bears great practical, as well as conceptual, interest.

Prior research in this direction includes the work of Csiszár [2], [3] who demonstrated that the error exponent of joint source-channel coding outperforms that of separate source-channel coding. For discrete source-channel pairs with average distortion criterion, Pilc's achievability bound [4], [5] applies. For the transmission of a Gaussian source over a discrete channel under the average mean square error constraint, Wyner's achievability bound [6], [7] applies. Nonasymptotic achievability and converse bounds for a graph-theoretic model of JSCC have been obtained by Csiszár [8]. Most recently, Tauste Campo et al. [9] showed a number of finite-blocklength random-coding bounds applicable to the almost-lossless JSCC setup, while Wang et al. [10] found the dispersion of JSCC for sources and channels with finite alphabets.

In this paper we give a non-asymptotic analysis of joint source-channel coding including several achievability and converse bounds, which hold in wide generality and are tight enough to determine the dispersion of joint source-channel coding for the transmission of an abstract memoryless source over either discrete memoryless channel (DMC) or a Gaussian channel, under an arbitrary fidelity criterion. We also investigate the penalty incurred by separate source-channel coding using both the source-channel dispersion and the particularization of our new bounds to (i) the binary source and the binary symmetric channel with bit error rate fidelity criterion and (ii) the Gaussian source and Gaussian channel under mean-square error distortion.

Further, we revisit the dilemma of whether one should or should not code when operating under delay constraints. Gastpar et al. [11] gave a set of necessary and sufficient conditions on the source, its distortion measure, the channel and its cost function in order for symbol-by-symbol transmission to attain the minimum average distortion. In these curious cases, the

source and the channel are probabilistically matched. We show that whenever the source and the channel are probabilistically matched so that symbol-by-symbol coding achieves the minimum average distortion, it also achieves the dispersion of joint source-channel coding. Moreover, even in the absence of such a match between the source and the channel, symbol-by-symbol transmission, though asymptotically suboptimal, might outperform in the non-asymptotic regime not only separate source-channel coding but also our random-coding achievability bound.

The non-asymptotic theoretical limit of interest in this paper is the maximum number of source symbols per channel input transmissible at a given channel blocklength under the fidelity constraint of exceeding a given distortion level, regardless of decoding complexity. The excess distortion constraint is, in a way, more fundamental than the average distortion constraint which is the figure of merit in [4]–[7], because it gives full information about the distribution (and not just its mean) of the distortion incurred at the decoder output.

The rest of the paper is organized as follows. Section II summarizes basic definitions and notation. Sections III and IV introduce the new converse and achievability bounds to the maximum achievable coding rate, respectively. A Gaussian approximation analysis of the new bounds is presented in Section V. The evaluation of the bounds and the approximation is performed for two important special cases: the transmission of a binary memoryless source (BMS) over a binary symmetric channel (BSC) with bit error rate distortion (Section VI) and the transmission of a Gaussian memoryless source (GMS) with mean-square error distortion over an AWGN channel with a total power constraint (Section VII). Section VIII focuses on symbol-by-symbol transmission.

II. Definitions

A lossy source-channel code is a pair of (possibly randomized) mappings $\mathsf{f}\colon \mathcal{M} \mapsto \mathcal{X}$ and $\mathsf{g}\colon \mathcal{Y} \mapsto \hat{\mathcal{M}}$. A distortion measure $\mathsf{d}\colon \mathcal{M} \times \hat{\mathcal{M}} \mapsto [0, +\infty)$ is used to quantify the performance of the lossy code. A cost function $\mathsf{c}\colon \mathcal{X} \mapsto [0, +\infty)$ may be imposed on the channel inputs.

Definition 1. *A (d, ϵ, α) code for $\{\mathcal{M},\ \mathcal{X},\ \mathcal{Y},\ \hat{\mathcal{M}},\ P_S, \mathsf{d},\ P_{Y|X},\ \mathsf{c}\}$ is a source-channel code with $\mathbb{P}\left[\mathsf{d}\left(S, \mathsf{g}(Y)\right) > d\right] \le \epsilon$ and either $\mathbb{E}\left[\mathsf{c}(X)\right] \le \alpha$ (average cost constraint) or $\sup_{x \in \mathcal{X}} \mathsf{c}(x) \le \alpha$ (maximal cost constraint), where $\mathsf{f}(S) = X$ (see Fig. 1)*

If there is no cost constraint ($\mathsf{c}(x) = 0$ for all $x \in \mathcal{X}$), we say a '(d, ϵ) code' instead of a

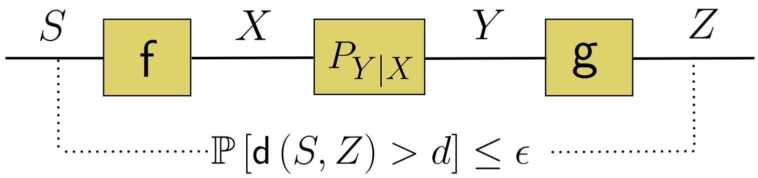


Fig. 1. A (d, ϵ) joint source-channel code.

'(d, ϵ, α) code'.

The special case $d = 0$ and $\mathsf{d}(s, z) = 1\{s \neq z\}$ corresponds to almost-lossless compression. If, in addition, P_S is equiprobable on an alphabet of cardinality $|\mathcal{X}| = |\mathcal{Y}| = M$, a $(0, \epsilon, \alpha)$ code in Definition 1 corresponds to an (M, ϵ, α) channel code (i.e. a code with M codewords and average error probability ϵ and cost α). On the other hand, if $P_{Y|X}$ is an identity mapping on an alphabet of cardinality M without cost constraints, a (d, ϵ) code in Definition 1 corresponds to an (M, d, ϵ) lossy compression code (as e.g. defined in [12]).

Definition 2. *In the conventional fixed-to-fixed (or block) setting in which $\mathcal{X}$ and $\mathcal{Y}$ are the $n-$fold Cartesian products of alphabets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{M}$ and $\hat{\mathcal{M}}$ are the $k-$fold Cartesian products of alphabets $\mathcal{S}$ and $\hat{\mathcal{S}}$, and $\mathsf{d}_k\colon \mathcal{S}^k \times \hat{\mathcal{S}}^k \mapsto [0, +\infty)$, $\mathsf{c}_n\colon \mathcal{A}^n \mapsto [0, +\infty)$, a (d, ϵ, α) code for $\{\mathcal{S}^k,\ \mathcal{A}^n,\ \mathcal{B}^n,\ \hat{\mathcal{S}}^k,\ P_{S^k},\ \mathsf{d}_k,\ P_{Y^n|X^n},\ \mathsf{c}_n\}$ is called a $(k, n, d, \epsilon, \alpha)$ code (a (k, n, d, ϵ) code if there is no cost constraint).*

Definition 3. *Fix ϵ, d, α and the channel blocklength n. The maximum achievable source blocklength and coding rate (source symbols per channel use) are defined by, respectively*

$$k^\star(n, d, \epsilon, \alpha) = \max\left\{k\colon \exists(k, n, d, \epsilon, \alpha) \text{ code}\right\} \tag{1}$$

$$R(n, d, \epsilon, \alpha) = \frac{k^\star(n, d, \epsilon, \alpha)}{n} \tag{2}$$

Alternatively, fix ϵ, α, source blocklength k and channel blocklength n. The minimum achievable excess distortion is defined by

$$D(k, n, \epsilon, \alpha) = \inf\left\{d\colon \exists(k, n, d, \epsilon, \alpha) \text{ code}\right\} \tag{3}$$

Denote, for a given $P_{Y|X}$,

$$\mathbb{C}(\alpha) = \sup_{\substack{P_X:\\ \mathbb{E}[\mathsf{c}(X)]\le\alpha}} I(X;Y) \tag{4}$$

and, for a given P_S and the distortion measure $\mathsf{d}\colon \mathcal{M}\times\hat{\mathcal{M}} \mapsto [0,+\infty)$,

$$\mathbb{R}_S(d) = \inf_{\substack{P_{Z|S}:\\ \mathbb{E}[\mathsf{d}(S,Z)]\le d}} I(S;Z) \tag{5}$$

We impose the following basic restrictions on $P_{Y|X}$, P_S and the distortion measure

(a) $\mathbb{R}_S(d)$ is finite for some d, i.e. $d_{\min} < \infty$, where

$$d_{\min} = \inf\left\{d\colon\ \mathbb{R}_S(d) < \infty\right\} \tag{6}$$

(b) The infimum in (5) is achieved by a unique $P_{Z^\star|S}$.

(c) The supremum in (4) is achieved by a unique $P_{X^\star}$.

The dispersion, which serves to quantify the penalty on the rate of the best JSCC code induced by the finite blocklength, is defined as follows.

Definition 4. *Fix α and $d \ge d_{\min}$. The rate-dispersion function of joint source-channel coding (source samples squared per channel use) is defined as*

$$\mathcal{V}(d,\alpha) = \lim_{\epsilon\to 0}\limsup_{n\to\infty} n\left(\frac{\frac{C(\alpha)}{R(d)} - R(n,d,\epsilon,\alpha)}{Q^{-1}(\epsilon)}\right)^2 \tag{7}$$

where $C(\alpha)$ and $R(d)$ are the channel capacity-cost and source rate-distortion functions, respectively.[1]

The distortion-dispersion function of joint source-channel coding is defined as

$$\mathcal{W}(R,\alpha) = \lim_{\epsilon\to 0}\limsup_{n\to\infty} n\left(\frac{D\left(\frac{C(\alpha)}{R}\right) - D(nR,n,\epsilon,\alpha)}{Q^{-1}(\epsilon)}\right)^2 \tag{8}$$

where $D(\cdot)$ is the distortion-rate function of the source.

If there is no cost constraint, we will simplify notation and drop α from (1), (2), (3), (4), (7) and (8).

[1]While for memoryless sources and channels, $C(\alpha) = \mathbb{C}(\alpha)$ and $R(d) = \mathbb{R}_S(d)$ given by (4) and (5) evaluated with single-letter distributions, it is important to distinguish between the operational definitions and the extremal mutual information quantities, since the core results in this paper allow for memory.

Definition 5 ($d-$tilted information [12]). *For $d > d_{\min}$, the $\mathsf{d}-$tilted information in s is defined as*[2]

$$\jmath_S(s,d) = \log \frac{1}{\mathbb{E}\left[\exp\left\{\lambda^\star d - \lambda^\star \mathsf{d}(s, Z^\star)\right\}\right]} \tag{9}$$

where the expectation is with respect to the unconditional distribution of $Z^\star$, and

$$\lambda^\star = -\mathbb{R}'_S(d) \tag{10}$$

The following properties of $\mathsf{d}-$tilted information, proven in [13], are used in the sequel.

$$\jmath_S(s,d) = \imath_{S;Z^\star}(s;z) + \lambda^\star \mathsf{d}(s,z) - \lambda^\star d \tag{11}$$

$$\mathbb{E}\left[\jmath_S(s,d)\right] = \mathbb{R}_S(d) \tag{12}$$

$$\mathbb{E}\left[\exp\left\{\lambda^\star d - \lambda^\star \mathsf{d}(S,z) + \jmath_S(S,d)\right\}\right] \le 1 \tag{13}$$

where (11) holds for $P_Z^\star$-almost every z, while (13) holds for all $z \in \hat{\mathcal{M}}$, and

$$\imath_{S;Z}(s;z) = \log \frac{dP_{Z|S=s}}{dP_Z}(z) \tag{14}$$

denotes the information density of the joint distribution P_{SZ} at (s,z). We can define the right side of (14) for a given $(P_{Z|S}, P_Z)$ even if there is no P_S such that the marginal of $P_S P_{Z|S}$ is P_Z. We use the same notation $\imath_{S;Z}$ for that more general function. To extend Definition 5 to the lossless case, for discrete random variables we define 0-tilted information as

$$\jmath_S(s,0) = \imath_S(s) \tag{15}$$

where

$$\imath_S(s) = \log \frac{1}{P_S(s)} \tag{16}$$

is the information in outcome $s \in \mathcal{M}$.

Finally, the distortion d-ball around $s \in \mathcal{M}$ is denoted by

$$B_d(s) = \{z \in \hat{\mathcal{M}} \colon \mathsf{d}(s,z) \le d\} \tag{17}$$

[2]All $\log$'s and $\exp$'s are in an arbitrary common base.

So as not to clutter notation, in Sections III and IV we assume that there are no cost constraints. However, all results in those sections generalize to the case of a maximal cost constraint by considering X whose distribution is supported in the subset of allowable channel inputs:

$$\mathcal{F}(\alpha) = \{x \in \mathcal{X} \colon \mathsf{c}(x) \leq \alpha\} \tag{18}$$

rather than the entire channel input alphabet $\mathcal{X}$.

III. Converses

A. Converses via d-tilted information

Our first result is a general converse bound.

Theorem 1 (Converse). *The existence of a (d, ϵ) code for S and $P_{Y|X}$ requires that*

$$\begin{aligned} \epsilon &\geq \inf_{P_{X|S}} \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \mathbb{P}\left[\jmath_S(S, d) - \imath_{X;\bar{Y}}(X; Y) \geq \gamma\right] - \exp(-\gamma) \right\} & (19) \\ &\geq \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \mathbb{E}\left[\inf_{x \in \mathcal{X}} \mathbb{P}\left[\jmath_S(S, d) - \imath_{X;\bar{Y}}(x; Y) \geq \gamma \mid S\right]\right] - \exp(-\gamma) \right\} & (20) \end{aligned}$$

where in (19)*, $S - X - Y$, and the conditional probability in* (20) *is with respect to Y distributed according to $P_{Y|X=x}$ (independent of S), and*

$$\imath_{X;\bar{Y}}(x; y) = \log \frac{dP_{Y|X=x}}{dP_{\bar{Y}}}(y) \tag{21}$$

Proof: Fix γ and the (d, ϵ, α) code $(P_{X|S}, P_{Z|X})$. Fix an arbitrary probability measure $P_{\bar{Y}}$ on $\mathcal{Y}$. Let $P_{\bar{Y}} \to P_{Z|Y} \to P_{\bar{Z}}$, i.e. $P_{\bar{Z}}(z) = \sum_{y \in \mathcal{Y}} P_{Z|Y}(z|y) P_{\bar{Y}}(y)$. [3] We can write the probability

[3]We write summations over alphabets for simplicity. Unless stated otherwise, all our results hold for abstract probability spaces.

in the right side of (19) as

$$\begin{aligned}
&\mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma\right] \\
&= \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma, \mathsf{d}(S;Z) > d\right] \\
&+ \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma, \mathsf{d}(S;Z) \leq d\right] && (22)\\
&\leq \epsilon + \sum_{s \in \mathcal{M}} P_S(s) \sum_{x \in \mathcal{X}} P_{X|S}(x|s) \sum_{y \in \mathcal{Y}} \sum_{z \in B_d(s)} P_{Z|Y}(z|y) \\
&\cdot P_{Y|X}(y|x) 1\left\{P_{Y|X}(y|x) \leq P_{\bar{Y}}(y) \exp\left(\jmath_S(s,d) - \gamma\right)\right\} && (23)\\
&\leq \epsilon + \exp\left(-\gamma\right) \sum_{s \in \mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y) \sum_{z \in B_d(s)} P_{Z|Y}(z|y) \sum_{x \in \mathcal{X}} P_{X|S}(x|s) && (24)\\
&= \epsilon + \exp\left(-\gamma\right) \sum_{s \in \mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y) \sum_{z \in B_d(s)} P_{Z|Y}(z|y) && (25)\\
&= \epsilon + \exp\left(-\gamma\right) \sum_{s \in \mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) P_{\bar{Z}}(B_d(s)) && (26)\\
&\leq \epsilon + \exp\left(-\gamma\right) \sum_{z \in \hat{\mathcal{M}}} P_{\bar{Z}}(z) \sum_{s \in \mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d) + \lambda^\star d - \lambda^\star \mathsf{d}(s,z)\right) && (27)\\
&\leq \epsilon + \exp\left(-\gamma\right) && (28)
\end{aligned}$$

where (28) is due to (13). Optimizing over $\gamma > 0$ and $P_{\bar{Y}}$ to obtain the best possible bound for a given encoder $P_{X|S}$. To obtain a code-independent converse, we simply optimize over $P_{X|S}$, and (19) follows. To show (20), we weaken (19) as

$$\epsilon \geq \sup_{\gamma > 0} \left\{ \sup_{P_{\bar{Y}}} \inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma\right] - \exp\left(-\gamma\right) \right\} \tag{29}$$

and observe that for any $P_{\bar{Y}}$,

$$\begin{aligned}
&\inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(X;Y) \geq \gamma\right] \\
&= \sum_{s \in \mathcal{M}} P_S(s) \inf_{P_{X|S=s}} \sum_{x \in \mathcal{X}} P_{X|S}(x|s) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) 1\left\{\jmath_S(s,d) - \imath_{X;\bar{Y}}(x;y) \geq \gamma\right\} && (30)\\
&= \sum_{s \in \mathcal{M}} P_S(s) \inf_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) 1\left\{\jmath_S(s,d) - \imath_{X;\bar{Y}}(x;y) \geq \gamma\right\} && (31)\\
&= \mathbb{E}\left[\inf_{x \in \mathcal{X}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar{Y}}(x;Y) \geq \gamma \mid S\right]\right] && (32)
\end{aligned}$$

■

An immediate corollary to Theorem 1 is the following result.

Theorem 2 (Converse)**.** *Assume that there exists a distribution $P_{\bar Y}$ such that the distribution of $\imath_{X;\bar Y}(x;Y)$ (according to $P_{Y|X=x}$) does not depend on the choice of $x \in \mathcal{X}$. If a (d,ϵ) code for S and $P_{Y|X}$ exists, then*

$$\epsilon \geq \sup_{\gamma>0} \left\{ \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;\bar Y}(x;Y) \geq \gamma\right] - \exp\left(-\gamma\right) \right\} \tag{33}$$

for an arbitrary $x \in \mathcal{X}$. The probability measure $\mathbb{P}$ in (33) *is generated by $P_S P_{Y|X=x}$.*

Proof: Observe that under the assumption the conditional probability in the right side of (20) is the same regardless of the choice of $x \in \mathcal{X}$. ∎

Remark 1. Theorems 1 and 2 still hold in the case $d = 0$ and $d(x,y) = 1\left\{x \neq y\right\}$ that corresponds to almost-lossless data compression. Indeed, recalling (15), it is easy to see that the proof of Theorem 1 applies, skipping the now unnecessary step (22), and, therefore, (19) reduces to

$$\epsilon \geq \inf_{P_{X|S}} \sup_{\gamma>0} \left\{ \sup_{P_{\bar Y}} \mathbb{P}\left[\imath_S(S) - \imath_{X;\bar Y}(X;Y) \geq \gamma\right] - \exp\left(-\gamma\right) \right\} \tag{34}$$

Remark 2. Our converse for lossy source coding in [12, Theorem 7] can be viewed as a particular case of the result in Theorem 2. Indeed, if $\mathcal{X} = \mathcal{Y} = \{1, \ldots, M\}$ and $P_{Y|X}(m|m) = 1$, $P_Y(1) = \ldots = P_Y(M) = \frac{1}{M}$, then (33) becomes

$$\epsilon \geq \sup_{\gamma>0} \mathbb{P}\left[\jmath_S(S,d) \geq \log M + \gamma\right] - \exp\left(-\gamma\right) \tag{35}$$

which is precisely [12, Theorem 7].

The next result generalizes Theorem 1. When we apply Theorem 3 in Section V to find the dispersion of JSCC, we will let T be the number of channel input types. If $T = 1$, Theorem 3 reduces to Theorem 1.

Theorem 3 (Converse)**.** *Fix a positive integer T and a partition $\{\mathcal{X}_t, t = 1, \ldots, T\}$ of $\mathcal{X}$. Define $\mathsf{T}\colon \mathcal{X} \mapsto [1, \ldots, T]$, $\mathsf{T}(x) = t$ if $x \in \mathcal{X}_t$. The existence of a (d,ϵ) code for S and $P_{Y|X}$ requires that*

$$\epsilon \geq \inf_{P_{X|S}} \sup_{\gamma>0} \left\{ \sup_{\left\{P_{Y_t}\right\}_{t=1}^T} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;Y_{\mathsf{T}(X)}}(X;Y) \geq \gamma\right] - T\exp\left(-\gamma\right) \right\} \tag{36}$$

$$\geq \sup_{\gamma>0} \left\{ \sup_{\left\{P_{Y_t}\right\}_{t=1}^T} \mathbb{E}\left[\inf_{x\in\mathcal{X}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;Y_{\mathsf{T}(x)}}(x;Y) \geq \gamma \mid S\right]\right] - T\exp\left(-\gamma\right) \right\} \tag{37}$$

where in (36)*,* $S - X - Y$*, and in* (37)*, the probability measure is generated by* $P_S P_{Y|X=x}$*,*

$$\imath_{X;Y_t}(x;y) = \log \frac{P_{Y|X=x}}{P_{Y_t}}(y) \tag{38}$$

and the supremum is over $\gamma > 0$ *and all collections* $\{P_{Y_1}, \ldots, P_{Y_T}\}$ *of probability distributions on the channel output alphabet* $\mathcal{Y}$ *parameterized by the partition index* t*.*

Proof: Fix an arbitrary collection of probability measures P_{Y_t} on $\mathcal{Y}$ parametrized by partition index t. Let $P_{Z_t}(z) = \sum_{y \in \mathcal{Y}} P_{Z|Y}(z|y) P_{Y_t}(y)$ where $P_{Z|Y}$ is the decoder.

$$\begin{aligned}
&\mathbb{P}\left[\jmath_S(S,d) - \imath_{X;Y_{\mathsf{T}(X)}}(X;Y) \geq \gamma\right] \\
&\leq \epsilon + \sum_{s \in \mathcal{M}} P_S(s) \sum_{t=1}^{T} \sum_{x \in \mathcal{X}_t} P_{X|S}(x|s) \sum_{y \in \mathcal{Y}} \sum_{z \in B_d(s)} P_{Z|Y}(z|y) \\
&\quad \cdot P_{Y|X}(y|x) 1\left\{P_{Y|X}(y|x) \leq P_{Y_t}(y) \exp\left(\jmath_S(s,d) - \gamma\right)\right\} && (39) \\
&\leq \epsilon + \exp(-\gamma) \sum_{t=1}^{T} \sum_{s \in \mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) \sum_{y \in \mathcal{Y}} P_{Y_t}(y) \sum_{z \in B_d(s)} P_{Z|Y}(z|y) \sum_{x \in \mathcal{X}_t} P_{X|S}(x|s) && (40) \\
&\leq \epsilon + \exp(-\gamma) \sum_{t=1}^{T} \sum_{s \in \mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) \sum_{y \in \mathcal{Y}} P_{Y_t}(y) \sum_{z \in B_d(s)} P_{Z|Y}(z|y) && (41) \\
&\leq \epsilon + \exp(-\gamma) \sum_{t=1}^{T} \sum_{s \in \mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d)\right) P_{Z_t}(B_d(s)) && (42) \\
&\leq \epsilon + \exp(-\gamma) \sum_{t=1}^{T} \sum_{z \in \hat{\mathcal{M}}} P_{Z_t}(z) \sum_{s \in \mathcal{M}} P_S(s) \exp\left(\jmath_S(s,d) + \lambda^\star d - \lambda^\star \mathsf{d}(s,z)\right) && (43) \\
&\leq \epsilon + T \exp(-\gamma) && (44)
\end{aligned}$$

where (44) is due to (13). Optimizing over $\gamma > 0$ and $P_{Y_1}, \ldots, P_{Y_T}$ to obtain the best possible bound for a given encoder $P_{X|S}$. To obtain a code-independent converse, we simply optimize over $P_{X|S}$, and (36) follows. To show (37), we weaken (36) as,

$$\epsilon \geq \sup_{\gamma > 0} \left\{ \sup_{\{P_{Y_t}\}_{t=1}^{T}} \inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;Y_{\mathsf{T}(X)}}(X;Y) \geq \gamma\right] - \exp(-\gamma) \right\} \tag{45}$$

and observe that for any choice of $\{P_{Y_1}, \ldots, P_{Y_T}\}$,

$$\inf_{P_{X|S}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;Y_{\mathsf{T}(X)}}(X;Y) \geq \gamma\right] \tag{46}$$

$$= \sum_{s\in\mathcal{M}} P_S(s) \inf_{P_{X|S=s}} \sum_{x\in\mathcal{X}} P_{X|S}(x|s) \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) 1\left\{\jmath_S(s,d) - \imath_{X;Y_{\mathsf{T}(x)}}(x;y) \geq \gamma\right\} \tag{47}$$

$$= \sum_{s\in\mathcal{M}} P_S(s) \inf_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) 1\left\{\jmath_S(s,d) - \imath_{X;Y_{\mathsf{T}(x)}}(x;y) \geq \gamma\right\} \tag{48}$$

$$= \mathbb{E}\left[\inf_{x\in\mathcal{X}} \mathbb{P}\left[\jmath_S(S,d) - \imath_{X;Y_{\mathsf{T}(x)}}(x;Y) \geq \gamma \mid S\right]\right] \tag{49}$$

■

B. Converses via hypothesis testing and list decoding

To show a joint source-channel converse in [3], Csiszár used a list decoder, which outputs a list of L elements drawn from $\mathcal{M}$. While traditionally list decoding has only been considered in the context of finite alphabet sources, we generalize the setting to sources with abstract alphabets. In our setup, the encoder is the random transformation $P_{X|S}$, and the decoder is defined as follows.

Definition 6 (List decoder). *An (L, Q_S) list decoder is a random transformation $P_{\tilde{S}|Y}$, where $\tilde{S}$ takes values on Q_S-measurable sets with Q_S-measure not exceeding L:*

$$Q_S\left(\tilde{S}\right) \leq L \tag{50}$$

The error probability with this type of list decoding is the probability that the source output S is not on the decoder output list for Y:

$$1 - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \sum_{\tilde{s}\in\mathcal{M}^{(L)}} \sum_{s\in\tilde{s}} P_{\tilde{S}|Y}(\tilde{s}|y) P_{Y|X}(y|x) P_{X|S}(x|s) P_S(s) \tag{51}$$

where $\mathcal{M}^{(L)}$ consists of Q_S-measurable subsets of $\mathcal{M}$ with Q_S-measure not exceeding L.

Definition 7 (List code). *An (ϵ, L, Q_S) list code is a pair of random transformations $(P_{X|S}, P_{\tilde{S}|Y})$ such that* (50) *holds and the list error probability* (51) *does not exceed ϵ.*

Of course, letting $Q_S = U_S$, where U_S is the counting measure on $\mathcal{M}$, we recover the conventional list decoder definition. The almost-lossless JSCC setting ($d = 0$) in Definition 1 corresponds to $L = 1$, $Q_S = U_S$. If the source alphabet is the real line, it is reasonable to let Q_S be the Lebesgue measure.

Any converse for list coding implies a converse for lossy coding. To see this, observe that any (d, ϵ) lossy code can be converted to a list code with list error probability not exceeding ϵ by feeding the lossy decoder output to a function that outputs the list of all source outcomes s within distortion d from the output $z \in \hat{\mathcal{M}}$ of the original lossy decoder. Therefore, the set of all (d, ϵ) lossy codes is included in the set of all list codes with list error probability $\leq \epsilon$ and list size

$$L = \max_{z \in \hat{\mathcal{M}}} Q_S \left(\{s \colon \mathsf{d}(s, z) \leq d\}\right) \tag{52}$$

Denote by

$$\beta_\alpha(P, Q) = \min_{\substack{P_{W|X}\colon \\ \mathbb{P}[W=1] \geq \alpha}} \mathbb{Q}\left[W = 1\right] \tag{53}$$

the optimal performance achievable among all randomized tests $P_{W|X} \colon \mathcal{X} \to \{0, 1\}$ between probability distributions P and Q on $\mathcal{X}$ is denoted by (1 indicates that the test chooses P)[4]. In fact, Q need not be a probability measure, it just needs to be σ-finite in order for the Neyman-Pearson lemma and related results to hold.

The hypothesis testing converse for channel coding [14, Theorem 27] can be generalized to joint source-channel coding with list decoding as follows.

Theorem 4 (Converse)**.** *Fix P_S and $P_{Y|X}$, and let Q_S be a σ-finite measure. The existence of an (ϵ, L, Q_S) list code requires that*

$$\inf_{P_{X|S}} \sup_{P_{\bar{Y}}} \beta_{1-\epsilon}(P_S P_{X|S} P_{Y|X}, Q_S P_{X|S} P_{\bar{Y}}) \leq L \tag{54}$$

where the supremum is over all probability measures $P_{\bar{Y}}$ defined on the channel output alphabet $\mathcal{Y}$.

Proof: Fix Q_S, the encoder $P_{X|S}$, and an auxiliary σ-finite conditional measure $Q_{Y|XS}$. Consider the (not necessarily optimal) test for deciding between $P_{SXY} = P_S P_{X|S} P_{Y|X}$ and $Q_{SXY} = Q_S P_{X|S} Q_{Y|XS}$ which chooses P_{SXY} if S is on the decoder output list when Y is observed at the channel output.

[4]Throughout, P, Q denote distributions, whereas $\mathbb{P}$, $\mathbb{Q}$ are used for the corresponding probabilities of events on the underlying probability space.

According to $\mathbb{P}$, the probability measure generated by P_{SXY}, the probability that the test chooses P_{SXY} is given by

$$\mathbb{P}\left[S \in \tilde{S}\right] \geq 1 - \epsilon \tag{55}$$

Since $\mathbb{Q}\left[S \in \tilde{S}\right]$ is the measure of the event that the test chooses P_{SXY} when Q_{SXY} is true, and the optimal test cannot perform worse than the possibly suboptimal one that we selected, it follows that

$$\beta_{1-\epsilon}(P_S P_{X|S} P_{Y|X}, Q_S P_{X|S} Q_{Y|XS}) \leq \mathbb{Q}\left[S \in \tilde{S}\right] \tag{56}$$

Now, fix an arbitrary probability measure $P_{\bar{Y}}$ on $\mathcal{Y}$. Choosing $Q_{Y|XS} = P_{\bar{Y}}$, the inequality in (56) can be weakened as follows.

$$\begin{aligned}
\mathbb{Q}\left[S \in \tilde{S}\right] &= \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y) \sum_{\tilde{s} \in \tilde{S}} P_{\tilde{S}|Y}(\tilde{s}|y) \sum_{s \in \tilde{s}} Q_S(s) \sum_{x \in \mathcal{X}} P_{X|S}(x|s) && (57)\\
&= \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y) \sum_{\tilde{s} \in \tilde{S}} P_{\tilde{S}|Y}(\tilde{s}|y) \sum_{s \in \tilde{s}} Q_S(s) && (58)\\
&\leq \sum_{y \in \mathcal{Y}} P_{\bar{Y}}(y) \sum_{\tilde{s} \in \tilde{S}} P_{\tilde{S}|Y}(\tilde{s}|y) L && (59)\\
&= L && (60)
\end{aligned}$$

Optimizing the bound over $P_{\bar{Y}}$ and choosing $P_{X|S}$ that yields the weakest bound in order to obtain a code-independent converse, (54) follows. ■

Remark 3. Similar to how Wolfowitz's converse for channel coding can be obtained from the meta-converse for channel coding [14], the converse for almost-lossless joint source-channel coding in (34) can be obtained by appropriately weakening (54) with $L = 1$. Indeed, invoking [14]

$$\beta_\alpha(P, Q) \geq \frac{1}{\gamma}\left(\alpha - \mathbb{P}\left[\frac{dP}{dQ} > \gamma\right]\right) \tag{61}$$

and letting $Q_S = U_S$ in (54), where U_S is the counting measure on $\mathcal{M}$, we have

$$\begin{aligned}
1 &\geq \inf_{P_{X|S}} \sup_{P_{\bar{Y}}} \beta_{1-\epsilon}(P_S P_{X|S} P_{Y|X}, U_S P_{X|S} P_{\bar{Y}}) && (62)\\
&\geq \inf_{P_{X|S}} \sup_{P_{\bar{Y}}} \sup_{\gamma > 0} \frac{1}{\gamma}\left(1 - \epsilon - \mathbb{P}\left[\imath_{X;\bar{Y}}(X;Y) - \imath_S(S) > \log \gamma\right]\right) && (63)
\end{aligned}$$

which upon rearranging yields (34).

In general, computing the infimum in (54) is challenging. However, if the channel is symmetric (in a sense formalized in the next result), $\beta_{1-\epsilon}(P_S P_{X|S} P_{Y|X}, U_S P_{X|S} P_{\bar{Y}})$ is independent of $P_{X|S}$.

Theorem 5 (Converse)**.** *Fix a probability measure $P_{\bar{Y}}$. Assume that the distribution of $\imath_{X;\bar{Y}}(x;Y)$ does not depend on $x \in \mathcal{X}$ under either $P_{Y|X=x}$ or $P_{\bar{Y}}$. Then, the existence of an (ϵ, L, Q_S) list code requires that*

$$\beta_{1-\epsilon}(P_S P_{Y|X=x}, Q_S P_{\bar{Y}}) \leq L \tag{64}$$

where $x \in \mathcal{X}$ is arbitrary.

Proof: The outcome of the optimum binary hypothesis test between P and Q only depends on $\frac{dP}{dQ}$ [15] for a given observation. In particular, the optimum binary hypothesis test $W^\star$ for deciding between $P_S P_{X|S} P_{Y|X}$ and $Q_S P_{X|S} P_{\bar{Y}}$ satisfies

$$W^\star - (S, \imath_{X;\bar{Y}}(X;Y)) - (S, X, Y) \tag{65}$$

For all $s \in \mathcal{M}$, $x \in \mathcal{X}$, we have

$$\begin{aligned}
\mathbb{P}\left[W^\star = 1 | S = s, X = x\right] &= \mathbb{E}\left[\mathbb{P}\left[W^\star = 1 | X = x, S = s, Y, \imath_{X;\bar{Y}}(x;Y)\right]\right] &(66)\\
&= \mathbb{E}\left[\mathbb{P}\left[W^\star = 1 | S = s, \imath_{X;\bar{Y}}(x;Y)\right]\right] &(67)\\
&= \sum_{y \in \mathcal{X}} P_{Y|X}(y|x) P_{W^\star|S=s,\imath_{X;\bar{Y}}(x;Y)}(1|S = s, \imath_{X;\bar{Y}}(x;Y)) &(68)\\
&= \mathbb{P}\left[W^\star = 1 | S = s\right] &(69)\\
\mathbb{Q}\left[W^\star = 1 | S = s, X = x\right] &= \mathbb{Q}\left[W^\star = 1 | S = s\right] &(70)
\end{aligned}$$

where

- (67) is due to (65),
- (68) uses the Markov property $S - X - Y$,
- (69) follows from the symmetry assumption on the distribution of $\imath_{X;\bar{Y}}(x, Y)$,
- (70) is obtained similarly to (68).

Since (69), (70) imply that the optimal test achieves the same performance (that is, the same $\mathbb{P}\left[W^\star = 1\right]$ and $\mathbb{Q}\left[W^\star = 1\right]$) regardless of $P_{X|S}$, we choose $P_{X|S} = \delta(x)$ for some $x \in \mathcal{X}$ in the left side of (54) to obtain (64). ∎

Remark 4. In the case of finite channel input and output alphabets, the channel symmetry assumption of Theorem 5 holds, in particular, if the rows of the channel transition probability matrix are permutations of each other, and $P_{\bar{Y}^n}$ is the equiprobable distribution on the (n-dimensional) channel output alphabet, which, coincidentally, is also the capacity-achieving output distribution. For Gaussian channels with equal power constraint, which corresponds to $\mathcal{X} = \{x : |x|^2 = nP\}$, any spherically-symmetric $P_{\bar{Y}^n}$ satisfies the assumption of Theorem 5.

IV. Achievability

Given a source code $(\mathsf{f}_{\mathrm{s}}^{(M)}, \mathsf{g}_{\mathrm{s}}^{(M)})$ of size M, and a channel code $(\mathsf{f}_{\mathrm{c}}^{(M)}, \mathsf{g}_{\mathrm{c}}^{(M)})$ of size M, we may concatenate them to obtain the following sub-class of the source-channel codes introduced in Definition 1:

Definition 8. *An (M, d, ϵ) source-channel code is a (d, ϵ) source-channel code such that the encoder and decoder mappings satisfy*

$$\mathsf{f} = \mathsf{f}_{\mathrm{c}}^{(M)} \circ \mathsf{f}_{\mathrm{s}}^{(M)} \tag{71}$$

$$\mathsf{g} = \mathsf{g}_{\mathrm{c}}^{(M)} \circ \mathsf{g}_{\mathrm{s}}^{(M)} \tag{72}$$

where

$$\mathsf{f}_{\mathrm{s}}^{(M)} \colon \mathcal{M} \mapsto \{1, \ldots, M\} \tag{73}$$

$$\mathsf{f}_{\mathrm{c}}^{(M)} \colon \{1, \ldots, M\} \mapsto \mathcal{X} \tag{74}$$

$$\mathsf{g}_{\mathrm{c}}^{(M)} \colon \mathcal{Y} \mapsto \{1, \ldots, M\} \tag{75}$$

$$\mathsf{g}_{\mathrm{s}}^{(M)} \colon \{1, \ldots, M\} \mapsto \hat{\mathcal{M}} \tag{76}$$

(see Fig. 2).

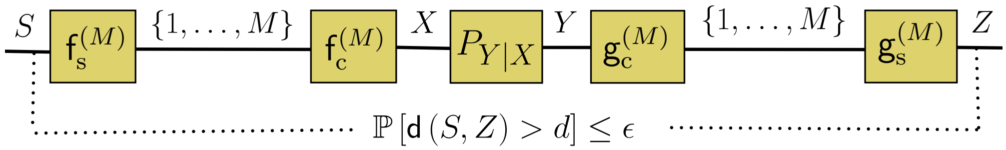


Fig. 2. A (d, ϵ) joint source-channel code.

The conventional separate source-channel coding paradigm corresponds to the special case of Definition 8 in which the source code $(\mathsf{f}_\mathsf{s}^{(M)}, \mathsf{g}_\mathsf{s}^{(M)})$ is chosen without knowledge of $P_{Y|X}$ and the channel code $(\mathsf{f}_\mathsf{s}^{(M)}, \mathsf{g}_\mathsf{s}^{(M)})$ is chosen without knowledge of P_S and the distortion measure d. If, in fact, both the source and the channel code are chosen optimally (i.e. with minimal distortion and error probability, respectively) for their given sizes, the separation principle guarantees that under certain conditions (which encompass the memoryless setting in this paper, see [16]) the asymptotic fundamental limit of joint source-channel coding is achievable. In the finite blocklength regime, however, such SSCC construction is, in general, only suboptimal. Within the SSCC paradigm, we can obtain an achievability result by further optimizing with respect to the choice of M:

Theorem 6 (Achievability, SSCC)**.** *Fix $P_{Y|X}$, d and P_S. Denote by $\epsilon^\star(M)$ the minimum achievable maximal error probability among all transmission codes of size M, and the minimum achievable probability of exceeding distortion d with a source code of size M by $\epsilon^\star(M, d)$.*

Then, there exists a (d, ϵ) source-channel code with

$$\epsilon \leq \min_{M} \{\epsilon^\star(M) + \epsilon^\star(M, d)\} \tag{77}$$

Bounds on $\epsilon^\star(M)$ and $\epsilon^\star(M, d)$ have been obtained recently in [14] and [12], respectively.

Definition 8 does not rule out choosing the source code based on the knowledge of $P_{Y|X}$ or the channel code based on the knowledge of P_S, d and d. One of the interesting conclusions in the present paper is that the optimal dispersion of JSCC is achievable within the class of (M, d, ϵ) source-channel codes introduced in Definition 8. However, the dispersion achieved by the conventional SSCC approach is in fact suboptimal.

To shed light on the reason behind the suboptimality of SSCC at finite blocklength despite its asymptotic optimality, consider a source code that, most of the time, encodes the source output within distortion $d > D\left(\frac{nC}{k}\right)$, for a fixed ratio $\frac{k}{n}$. Recall that the reason SSCC achieves the asymptotic fundamental limit is that the output of the source encoder is, for large k, approximately equiprobable over a set of roughly $\exp\left(kR(d)\right)$ distinct messages. From the channel coding theorem we know that there exists a channel code with the maximum likelihood decoder that is capable of distinguishing, with high probability, $M = \exp\left(kR(d)\right) < \exp\left(nC\right)$ messages. Therefore, simply putting the source code and the channel code together asymptotically results

in the maximum distortion d. Since d can be chosen to be arbitrarily close to its optimum value $D\left(\frac{nC}{k}\right)$, such separated scheme is asymptotically optimal.

However, at finite n, the output of the optimum source encoder is not, in general, equiprobable or nearly equiprobable, so there is no reason to expect that a separated scheme employing a channel code equipped with the maximum-likelihood detector would achieve the optimum nonasymptotic performance. Indeed, in the non-asymptotic regime the gain afforded by taking into account the residual encoded source redundancy at the channel decoder is appreciable. The following achievability result, obtained using independent random source codes and random channel codes within the paradigm of Definition 8, capitalizes on this intuition.

Theorem 7 (Achievability)**.** *For every positive integer M, there exists an (M, d, ϵ) source-channel code with*

$$\epsilon \leq \inf_{P_X, P_Z, \gamma > 0} \left\{ \mathbb{E}\left[\exp\left\{ - \left| \imath_{X;Y}(X;Y) - \log \frac{\gamma H_M}{P_Z(B_d(S))} \right|^+ \right\} \right] \right. \\ \left. + \mathbb{E}\left[\left| e^{-\gamma} - (1 - P_Z(B_d(S)))^M \right|^+ \right] + \mathbb{E}\left[(1 - P_Z(B_d(S)))^M \right] \right\} \tag{78}$$

where the expectations are with respect to $P_S P_X P_{Y|X} P_Z$ defined on $\mathcal{M} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{M}}$, and

$$H_M = \sum_{m=1}^{M} \frac{1}{m} \tag{79}$$

is the $M-$th harmonic number.

Proof: We construct a code with separate encoders for source and channel and separate decoders for source and channel as in Definition 8. We will perform a random coding analysis by choosing random independent source and channel codes which will lead to the conclusion that a pair of source and channel codes results in the error probability guaranteed in (78).

Source Encoder. Given an ordered list of representation points $(z_1, \ldots, z_M) \in \hat{\mathcal{M}}^M$, the source encoder selects the lowest index $m \in \{1, \ldots, M\}$ such that the source outcome is within distance d of z_m. If no such index can be found, the source encoder outputs an arbitrary index, e.g. M. Therefore,

$$\mathsf{f}_{\mathsf{s}}^{(M)}(s) = \begin{cases} m & \mathsf{d}(s, z_m) \leq d < \min_{i=1,\ldots,m-1} \mathsf{d}(s, z_i) \\ M & d < \min_{i=1,\ldots,M-1} \mathsf{d}(s, z_i) \end{cases} \tag{80}$$

In a good (M, d, ϵ) JSCC code, M would be chosen so large that with overwhelming probability, a source outcome would be encoded successfully within distortion d.

Channel Encoder. Given a codebook $(x_1, \ldots, x_M) \in \mathcal{X}^M$ the channel encoder outputs x_m if m is the output of the source encoder:

$$\mathsf{f}_{\mathrm{c}}^{(M)}(m) = x_m \tag{81}$$

Channel Decoder. Having observed $y \in \mathcal{Y}$, the channel decoder chooses arbitrarily among the members of the set:

$$\mathsf{g}_{\mathrm{c}}^{(M)}(y) = m \in \arg \max_{j \in \{1, \ldots M\}} \frac{P_{Y|X}(y|x_j)}{j} \tag{82}$$

A MAP decoder would multiply $P_{Y|X}(y|x_j)$ by $P_X(x_j)$. While that decoder would be too hard to analyze, the ratio in (82) is a good approximation because, averaging over all source codebooks, the probability that the j-th index is chosen is approximately proportional to $\frac{1}{j}$ (e.g. [17, (3)]).

Source Decoder. The source decoder outputs z_m if m is the output of the channel decoder:

$$\mathsf{g}_{\mathrm{s}}^{(M)}(m) = z_m \tag{83}$$

Error Probability Analysis. We now proceed to analyze the performance of the code described above. A channel decoding error can occur if and only if

$$\exists j \neq m \colon \; \frac{P_{Y|X}(Y|x_j)}{j} \geq \frac{P_{Y|X}(Y|x_m)}{m} \tag{84}$$

Let the channel codebook $(X_1, \ldots, X_M)$ be drawn i.i.d. from P_X, and independent of the source codebook $(Z_1, \ldots, Z_M)$, which is drawn i.i.d. from P_Z. Denote by $\epsilon(x^M, z^M)$ the excess-distortion probability attained with the source codebook z^M and the channel codebook x^M. Define the random variable $U \in \{1, \ldots, M+1\}$ which is a function of S and z^M only:

$$U = \begin{cases} \mathsf{f}_{\mathrm{s}}^{(M)}(S) & \mathsf{d}(S, \mathsf{g}_{\mathrm{s}}(\mathsf{f}_{\mathrm{s}}(S)) \leq d \\ M+1 & \text{otherwise} \end{cases} \tag{85}$$

Conditioned on the event $\{\mathsf{d}(S, \mathsf{g}_{\mathrm{s}}(\mathsf{f}_{\mathrm{s}}(S)) \leq d\}$ (no failure at the source encoder), the probability of excess distortion is upper bounded by the probability that the channel decoder does not choose $\mathsf{f}_{\mathrm{s}}^{(M)}(S)$, so

$$\begin{aligned} \epsilon(x^M, z^M) \leq \sum_{m=1}^{M} P_{U|Z^M}(m|z^M)\, \mathbb{P}\left[\bigcup_{j \neq m} \left\{ \frac{m P_{Y|X}(Y|x_j)}{j P_{Y|X}(Y|x_m)} \geq 1 \right\} \mid X = x_m \right] \\ + P_{U|Z^M}(M+1|z^M) \end{aligned} \tag{86}$$

We now average (86) over the source and channel codebooks. Averaging the m-th term of the sum in (86) with respect to the channel codebook yields

$$P_{U|Z^M}(m|z^M)\,\mathbb{P}\left[\bigcup_{j\neq m}\left\{\frac{mP_{Y|X}(Y|X_j)}{jP_{Y|X}(Y|X_m)}\geq 1\right\}\right] \tag{87}$$

where $Y, X_1, \ldots, X_M$ are distributed according to

$$P_{YX_1\ldots X_m}(y, x_1, \ldots, x_M) = P_{Y|X_m}(y|x_m)\prod_{j\neq m} P_X(x_j) \tag{88}$$

Letting $\bar{X}$ be an independent copy of X and applying the union bound to the probability in (87), we have

$$\begin{aligned}\mathbb{P}\left[\bigcup_{j\neq m}\left\{\frac{mP_{Y|X}(Y|X_j)}{jP_{Y|X}(Y|X_m)}\geq 1\right\}\right] &\leq \mathbb{E}\left[\min\left\{1, \sum_{j=1}^{M}\mathbb{P}\left[\frac{mP_{Y|X}(Y|\bar{X})}{jP_{Y|X}(Y|X)}\geq 1|X,Y\right]\right\}\right] &(89)\\ &\leq \mathbb{E}\left[\min\left\{1, \sum_{j=1}^{M}\frac{m\mathbb{E}\left[P_{Y|X}(Y|\bar{X})|Y\right]}{jP_{Y|X}(Y|X)}\right\}\right] &(90)\\ &= \mathbb{E}\left[\min\left\{1, H_M\frac{mP_Y(Y)}{P_{Y|X}(Y|X)}\right\}\right] &(91)\\ &= \mathbb{E}\left[\exp\left(-|\imath_{X;Y}(X;Y) - \log m - \log H_M|^+\right)\right] &(92)\end{aligned}$$

where

- (90) is due to $1\{a \geq 1\} \leq a$;
- (92) is due to $\min\{1, a\} = \exp\left(-\left|\log\frac{1}{a}\right|^+\right)$, where a is nonnegative.

On the other hand, conditioned on $S = s$ and averaged over the source codebook, U is distributed as:

$$P_{U|S}(m|s) = \begin{cases}\rho(s)(1-\rho(s))^{m-1} & m = 1, 2, \ldots, M\\ (1-\rho(s))^M & m = M+1\end{cases} \tag{93}$$

where we denoted for brevity

$$\rho(s) = P_Z(B_d(s)) \tag{94}$$

Averaging (86) further over the source codebook and using (92) and (93), we conclude

$$\mathbb{E}\left[\epsilon(Z^M, X^M)\right] \leq \inf_{P_X, P_Z}\mathbb{E}\left[\exp -|\imath_{X;Y}(X;Y) - \log U - \log H_M|^+ 1\{U\leq M\}\right] + \mathbb{E}\left[(1-\rho(S))^M\right] \tag{95}$$

Finally, we further upper bound the first expectation in the right side of (95) to make it more easily computable and analyzable. To that end, note that regardless of whether $j \leq M$, (93) results in

$$\mathbb{P}\left[j < U \leq M \mid S = s\right] = \left|(1-\rho(s))^k - (1-\rho(s))^M\right|^+ \tag{96}$$

Fix an arbitrary $\gamma > 0$ and observe that

$$(1-\rho(s))^{\left\lfloor \frac{\gamma}{\rho(s)} \right\rfloor} \leq (1-\rho(s))^{\frac{\gamma}{\rho(s)}-1} \tag{97}$$

$$\leq e^{-\rho(s)\left(\frac{\gamma}{\rho(s)}-1\right)} \tag{98}$$

$$\leq e^{-\gamma} \tag{99}$$

So, letting $j = \left\lfloor \frac{\gamma}{\rho(s)} \right\rfloor$, we obtain from (96) and (99) that

$$\mathbb{P}\left[\left\lfloor \frac{\gamma}{\rho(s)} \right\rfloor < U \leq M \mid S = s\right] \leq \left|e^{-\gamma} - (1-\rho(s))^M\right|^+ \tag{100}$$

Using

$$1\{U \leq M\} \leq 1\left\{U \leq \frac{\gamma}{\rho(S)}\right\} + 1\left\{\left\lfloor \frac{\gamma}{\rho(S)} \right\rfloor < U \leq M\right\} \tag{101}$$

and (100), we upper bound the first term in (95) as

$$\mathbb{E}\left[\exp -|\imath_{X;Y}(X;Y) - \log U - \log H_M|^+ 1\{U \leq M\}\right] \tag{102}$$

$$\leq \mathbb{E}\left[\exp -|\imath_{X;Y}(X;Y) - \log U - \log H_M|^+ 1\left\{U \leq \frac{\gamma}{\rho(S)}\right\}\right] + \mathbb{P}\left[\left\lfloor \frac{\gamma}{\rho(S)} \right\rfloor < U \leq M\right] \tag{103}$$

$$\leq \mathbb{E}\left[\exp -\left|\imath_{X;Y}(X;Y) - \log \frac{\gamma}{\rho(S)} - \log H_M\right|^+\right] + \mathbb{E}\left[\left|e^{-\gamma} - (1-\rho(S))^M\right|^+\right] \tag{104}$$

Finally, (78) follows by weakening (95) by means of (104) and invoking Shannon's random coding argument. ■

The code size M that leads to tight achievability bounds following from Theorem 7 is in general quite different from the size that achieves the minimum in (77). In that case, M is chosen so that $\log M$ lies between $kR(d)$ and nC so as to minimize the sum of source and channel decoding error probabilities without the benefit of a channel decoder that exploits residual source redundancy. In contrast, Theorem 7 is obtained with an approximate MAP decoder that allows a larger choice for $\log M$, even beyond nC. Still we can achieve a good (d, ϵ) tradeoff because the

channel code employs unequal error protection: those codewords with lower indices are more reliably decoded.

In the case of almost-lossless JSCC, the bound in Theorem 7 can be sharpened as shown recently by Tauste Campo et al. [9].

Theorem 8 (Achievability, almost-lossless JSCC [9])**.** *There exists a* $(0, \epsilon)$ *code with*

$$\epsilon \leq \inf_{P_X} \mathbb{E}\left[\exp\left(-|\imath_{X;Y}(X;Y) - \imath_S(S)|^+\right)\right] \tag{105}$$

where the expectation is with respect to $P_S P_X P_{Y|X}$ *defined on* $\mathcal{M} \times \mathcal{X} \times \mathcal{Y}$*.*

V. Gaussian Approximation

In addition to the basic conditions (a)-(c) of Section II, in this section we impose the following restrictions.

(i) The channel is stationary and memoryless, $P_{Y^n|X^n} = P_{\mathsf{Y}|\mathsf{X}} \times \ldots \times P_{\mathsf{Y}|\mathsf{X}}$.

(ii) The source is stationary and memoryless, $P_{S^k} = P_{\mathsf{S}} \times \ldots \times P_{\mathsf{S}}$, and the distortion measure is separable, $\mathsf{d}_k(s^k, z^k) = \frac{1}{k}\sum_{i=1}^k \mathsf{d}(s_i, z_i)$.

(iii) The distortion level satisfies $d_{\min} < d < d_{\max}$, where $d_{\min}$ is defined in (6), and $d_{\max} = \inf_{\mathsf{z} \in \hat{\mathcal{C}}} \mathbb{E}\left[\mathsf{d}(\mathsf{S}, \mathsf{z})\right]$, where the average is with respect to the unconditional distribution of S. The excess-distortion probability satisfies $0 < \epsilon < 1$.

(iv) $\mathbb{E}\left[\mathsf{d}^9(\mathsf{S}, \mathsf{Z}^\star)\right] < \infty$ where the average is with respect to $P_{\mathsf{S}} \times P_{\mathsf{Z}^\star}$ and $P_{\mathsf{Z}^\star}$ is the output distribution corresponding to the minimizer in (5).

Conditions (i) and (ii) are standard in the memoryless joint source-channel coding problem setup. The technical condition (iv) ensures applicability of the Gaussian approximation in Theorem 9 below.

Theorem 9 (Gaussian approximation)**.** *Under restrictions* (i)–(iv)*, the parameters of the optimal* (k, n, d, ϵ) *code satisfy*

$$nC - kR(d) = \sqrt{nV + k\mathcal{V}(d)}Q^{-1}(\epsilon) + \theta(n) \tag{106}$$

where

$$\mathcal{V}(d) = \mathrm{Var}\left[\jmath_{\mathsf{S}}(\mathsf{S}, d)\right] \tag{107}$$

and $k = O(n)$*,*

1) If $\mathcal{A}$ and $\mathcal{B}$ are finite and the channel has no cost constraints,

$$V = \mathrm{Var}\left[\imath^{\star}_{\mathsf{X};\mathsf{Y}}(\mathsf{X}^{\star};\mathsf{Y}^{\star})\right] \tag{108}$$

$$\imath^{\star}_{\mathsf{X};\mathsf{Y}}(x;y) = \log \frac{dP_{\mathsf{Y}|\mathsf{X}=x}}{dP_{\mathsf{Y}^{\star}}}(y) \tag{109}$$

where $\mathsf{X}^{\star}$, $\mathsf{Y}^{\star}$ are the capacity-achieving input and output random variables.

2) If the channel is Gaussian with either equal or maximal power constraint,

$$V = \frac{1}{2}\left(1 - \frac{1}{(1+P)^2}\right)\log^2 e \tag{110}$$

where P is the signal-to-noise ratio.

3) If $V > 0$,

$$-\underline{c}\log n + O\left(1\right) \le \theta\left(n\right) \tag{111}$$

$$\le \bar{c}\log n + \log\log n + O\left(1\right) \tag{112}$$

where

$$\underline{c} = |\mathcal{A}| - \frac{1}{2} \tag{113}$$

$$\bar{c} = 2 + \frac{\mathrm{Var}\left[\Lambda'_{\mathsf{Y}^{\star}}(\mathsf{X},\lambda^{\star})\right]}{\mathbb{E}\left[|\Lambda''_{\mathsf{Y}^{\star}}(\mathsf{X},\lambda^{\star})|\right]\log e} \tag{114}$$

In (114), *$(\cdot)'$ denotes differentiation with respect to λ, $\Lambda_{\mathsf{Y}^{\star}}(\mathsf{x},\lambda)$ is defined by*

$$\Lambda_{\mathsf{Y}^{\star}}(\mathsf{x},\lambda) = \log\frac{1}{\mathbb{E}\left[\exp\left(\lambda d - \lambda d(\mathsf{x},\mathsf{Z}^{\star})\right)\right]} \tag{115}$$

(cf. Definition 5) and $\lambda^{\star} = -R'(d)$.

4) If $V = 0$, (112) *still holds, while* (111) *is replaced with*

$$o\left(\sqrt{n}\right) \le \theta\left(n\right) \tag{116}$$

5) If the channel is such that the (conditional) distribution of $\imath^{\star}_{\mathsf{X};\mathsf{Y}}(\mathsf{x};\mathsf{Y})$ does not depend on $x \in \mathcal{X}$ or Gaussian with either equal or maximal power constraint, then $\underline{c} = \frac{1}{2}$.

6) In the almost-lossless case, $R(d) = H(\mathsf{S})$, and provided that the third absolute moment of $\imath_{\mathsf{S}}(\mathsf{S})$ is finite, (106) *and* (111) *still hold, while* (112) *strengthens to*

$$\theta\left(n\right) \le \frac{1}{2}\log n + O\left(1\right) \tag{117}$$

Proof: The asymptotic analysis of the bounds in Theorems 2 (converse, symmetric channel), 3 (converse, general lossy coding), 4 (converse, lossless coding) and 7 (achievability) is detailed in Appendices C and D. ∎

Remark 5. If the channel and the data compression codes are designed separately, we can invoke channel coding [14] and lossy compression [12] results to show that,

$$nC - kR(d) \leq \min_{\eta+\zeta\leq\epsilon}\left\{\sqrt{nV}Q^{-1}\left(\eta\right) + \sqrt{k\mathcal{V}(d)}Q^{-1}\left(\zeta\right)\right\} + O\left(\log(n+k)\right) \tag{118}$$

Comparing (118) to (106), observe that if either the channel or the source (or both) have zero dispersion, the joint source-channel coding dispersion can be achieved by separate coding. In that special case, either the d-tilted information or the channel density are so close to being deterministic that there is no need to account for the true distributions of these random variables, as a good joint source-channel code would do.

The Gaussian approximations of JSCC and SSCC in (106) and (118) admit the following heuristic interpretation when n is large and (thus, so is k). Since the source is stationary and memoryless, the normalized d-tilted information $J = \frac{1}{n}\jmath_{S^k}\left(S^k, d\right)$ becomes approximately Gaussian with mean $\frac{k}{n}R(d)$ and variance $\frac{k}{n}\frac{\mathcal{V}(d)}{n}$. Likewise, the normalized channel information density $I = \frac{1}{n}\imath_{X^n;Y^n}(X^n;Y^n)$ is, for large k, n, approximately Gaussian with mean C and variance $\frac{V}{n}$. Since the source is independent of the channel, the random variable $I - J$ is approximately Gaussian with mean $C - \frac{k}{n}R(d)$ and variance $\frac{1}{n}\left(\frac{k}{n}\mathcal{V}(d) + V\right)$, and (106) reflects the intuition that under JSCC, the source is reconstructed successfully within distortion d if and only if the channel information density exceeds the source d-tilted information, that is, $\{I > J\}$. In contrast, in SSCC, the source is reconstructed successfully if (I, J) falls into the intersection of half-planes $\{I > r\} \cap \{J < r\}$ for some $r = \frac{\log M}{n}$, which is the capacity of the noiseless link between the source and the channel code block that can be chosen so as to minimize the probability of that intersection, as reflected in (118). Since in JSCC the successful transmission event is strictly larger than in SSCC, i.e. $\{I > r\} \cap \{J < r\} \subset \{I > J\}$, separate source/channel code design incurs a performance loss. It is worth pointing out that $\{I > J\}$ leads to successful reconstruction even within the paradigm of the codes in Definition 8 because, as explained after

the proof of Theorem 7, unlike the SSCC case, it is not necessary that $\frac{\log M}{n}$ lie between I and J for successful reconstruction.

Remark 6. Using Theorem 9, it can be shown that

$$R(n,d,\epsilon,\alpha) = \frac{C(\alpha)}{R(d)} - \sqrt{\frac{\mathcal{V}(d)}{n}} Q^{-1}(\epsilon) - \frac{1}{R(d)}\frac{\theta(n)}{n} \tag{119}$$

where the rate-dispersion function of JSCC is found as (recall Definition 4),

$$\mathcal{V}(d) = \frac{R(d)V + C\mathcal{V}(d)}{R^3(d)} \tag{120}$$

Remark 7. Under regularity conditions similar to those in [12, Theorem 14], it can be shown that

$$D(nR,n,\epsilon,\alpha) = D\left(\frac{C(\alpha)}{R}\right) + \sqrt{\frac{\mathcal{W}(d)}{n}} Q^{-1}(\epsilon) + \frac{D'(R)}{R^2}\frac{\theta(n)}{n} \tag{121}$$

where the distortion-dispersion function of JSCC is given by

$$\mathcal{W}(R) = \left(\frac{D'(R)}{R}\right)^2 (V + R\mathcal{V}(D(C))) \tag{122}$$

Remark 8. If the basic conditions (b) and/or (c) fail so that there are several distributions $P_{\mathsf{Z}^\star|S}$ and/or several $P_{\mathsf{X}^\star}$ that achieve the rate-distortion function and the capacity, then

$$\mathcal{V}(d) \leq \min \mathcal{V}_{\mathsf{Z}^\star;\mathsf{X}^\star}(d) \tag{123}$$

$$\mathcal{W}(d) \leq \min \mathcal{W}_{\mathsf{Z}^\star;\mathsf{X}^\star}(d) \tag{124}$$

where the minimum is taken over $P_{\mathsf{Z}^\star|S}$ and $P_{\mathsf{X}^\star}$, and $\mathcal{V}_{\mathsf{Z}^\star;\mathsf{X}^\star}(d)$ (resp. $\mathcal{W}_{\mathsf{Z}^\star;\mathsf{X}^\star}(d)$) denotes (120) (resp. (122)) computed with $P_{\mathsf{Z}^\star|S}$ and $P_{\mathsf{X}^\star}$. The reason for possibly lower achievable dispersion in this case is that we have the freedom to map the unlikely source realizations leading to high probability of failure to those codewords resulting in the maximum variance so as to increase the probability that the channel output escapes the decoding failure region.

Remark 9. The dispersion of the Gaussian channel is given by (110), regardless of whether an equal or a maximal power constraint is imposed. An equal power constraint corresponds to the subset of allowable channel inputs being the power sphere:

$$\mathcal{F}(P) = \left\{ x^n \in \mathbb{R}^n \colon \frac{|x^n|}{\sigma_{\mathsf{N}}^2} = nP \right\} \tag{125}$$

where σ_{N}^2 is the noise power. In a maximal power constraint, (125) is relaxed replacing '$=$' with '$\leq$'.

Writing eq for the equal and $\max$ for the maximal power constraint, we remark that the bounds for the latter can be obtained from the bounds for the former via the following relation

$$k^\star_{\mathrm{eq}}(n,d,\epsilon) \le k^\star_{\max}(n,d,\epsilon) \le k^\star_{\mathrm{eq}}(n+1,d,\epsilon) \tag{126}$$

where the right-most inequality is due to the following idea dating back to Shannon: a (k,n,d,ϵ) code with a maximal power constraint can be converted to a $(k,n+1,d,\epsilon)$ code with an equal power constraint by appending an $(n+1)$-th coordinate to each codeword to equalize its total power to $n\sigma_{\mathsf{N}}^2 P$. From (126) it is immediate that the channel dispersions for maximal or equal power constraints must be the same.

VI. Lossy transmission of a BMS over a BSC

In this section we particularize the bounds in Sections III, IV and the approximation in Section V to the transmission of a BMS with bias p over a BSC with crossover probability δ. The target bit error rate $d \le p$.

The rate-distortion function of the source and the channel capacity are given by, respectively,

$$R(d) = h(p) - h(d) \tag{127}$$

$$C = 1 - h(\delta) \tag{128}$$

The source and the channel dispersions are given by [12], [14]:

$$\mathcal{V}(d) = p(1-p)\log^2 \frac{1-p}{p} \tag{129}$$

$$V = \delta(1-\delta)\log^2 \frac{1-\delta}{\delta} \tag{130}$$

where note that (129) does not depend on d.

Throughout the section, $w(a^\ell)$ denotes the Hamming weight of the binary ℓ-vector a^ℓ, and T^ℓ_α denotes a binomial random variable with parameters ℓ and α, independent of all other random variables. In addition, the binomial sum is denoted by

$$\left\langle {k \atop \ell} \right\rangle = \sum_{i=1}^{\ell} \binom{k}{i} \tag{131}$$

A straightforward particularization of the d-tilted information converse in Theorem 2 leads to the following result.

Theorem 10 (Converse, BMS-BSC)**.** *Any* (k, n, d, ϵ) *code for transmission of a BMS with bias* p *over a BSC with bias* δ *must satisfy*

$$\epsilon \geq \sup_{\gamma \geq 0} \left\{ \mathbb{P}\left[\left(T_p^k - kp \right) \log \frac{1-p}{p} + \left(T_\delta^n - n\delta \right) \log \frac{1-\delta}{\delta} \geq nC - kR(d) + \gamma \right] - \exp(-\gamma) \right\} \tag{132}$$

Note that the terms to the left of the '$\geq$' sign inside the probability in (132) are zero-mean random variables whose variances are equal to $k\mathcal{V}(d)$ and nV, respectively.

Proof: Let $P_{\bar{Y}^n} = P_{Y^{n\star}}$, which is the equiprobable distribution on $\{0,1\}^n$. An easy exercise reveals that

$$\jmath_{S^k}(s^k, d) = \imath_{S^k}(s^k) - kh(d) \tag{133}$$

$$\imath_{S^k}(s^k) = kh(p) + \left(w(s^k) - kp \right) \log \frac{1-p}{p} \tag{134}$$

$$\imath_{X^n;Y^{n\star}}(x^n; y^n) = n\left(\log 2 - h(\delta)\right) - \left(w(y^n - x^n) - n\delta\right) \log \frac{1-\delta}{\delta} \tag{135}$$

Since $w(Y^n - x^n)$ is distributed as T_δ^n regardless of $x^n \in \{0,1\}^n$, and $w(S^k)$ is distributed as T_p^k, Theorem 2 applies, and (33) becomes (132). ■

The hypothesis-testing converse in Theorem 4 particularizes to the following result:

Theorem 11 (Converse, BMS-BSC)**.** *Any* (k, n, d, ϵ) *code for transmission of a BMS with bias* p *over a BSC with bias* δ *must satisfy*

$$\mathbb{P}\left[U\left(\frac{1}{2}, \frac{1}{2}\right) < r \right] + \lambda \mathbb{P}\left[U\left(\frac{1}{2}, \frac{1}{2}\right) = r \right] \leq \left\langle \begin{matrix} k \\ \lfloor kd \rfloor \end{matrix} \right\rangle 2^{-k} \tag{136}$$

where the discrete random variable $U(\alpha, \beta)$ *is given by*

$$U(\alpha, \beta) = T_c^k \log \frac{1-p}{p} + T_\beta^n \log \frac{1-\delta}{\delta} \tag{137}$$

and $0 \leq \lambda < 1$ *and* r *are uniquely defined by*

$$\mathbb{P}\left[U(p, \delta) < r\right] + \lambda \mathbb{P}\left[U(p, \delta) = r\right] = 1 - \epsilon \tag{138}$$

Proof: As in the proof of Theorem 10, we let $P_{\bar{Y}^n}$ be the equiprobable distribution on $\{0,1\}^n$, $P_{\bar{Y}^n} = P_{Y^{n\star}}$. Since under $P_{Y^n|X^n = x^n}$, $w(Y^n - x^n)$ is distributed as T_δ^n, and under

$P_{Y^{n\star}}$, $w\left(Y^n - x^n\right)$ is distributed as $T^n_{\frac{1}{2}}$, irrespective of the choice of $x^n \in A^n$, according to both measures the distribution of the information density in (135) does not depend on the choice of x^n, so Theorem 5 applies. Further, we choose Q_{S^k} to be the equiprobable distribution on $\{0,1\}^k$ and observe that under P_{S^k}, the random variable $w(S^k)$ in (134) has the same distribution as T^k_p, while under Q_{S^k} it has the same distribution as $T^k_{\frac{1}{2}}$. Therefore, the log-likelihood ratio for testing between $P_{S^k}P_{Y^n|X^n=x^n}$ and $Q_{S^k}P_{Y^{n\star}}$ has the same distribution as ('$\sim$' denotes equality in distribution)

$$\log \frac{P_{S^k}(S^k)P_{Y^n|X^n=x^n}(Y^n)}{Q_{S^k}(S^k)P_{Y^{n\star}}(Y^n)} = \imath_{X^n;Y^n}(x^n;Y^n) - \imath_{S^k}(S^k) + k\log 2 \tag{139}$$

$$\sim n\log(2-2\delta) - k\log\frac{1}{1-p} - \begin{cases} U(p,\delta) & \text{under } P_{S^k}P_{Y^n|X^n=x^n} \\ U\left(\frac{1}{2},\frac{1}{2}\right) & \text{under } Q_{S^k}P_{Y^{n\star}} \end{cases} \tag{140}$$

so $\beta_{1-\epsilon}(P_{S^k}P_{Y^n|X^n=x^n}, Q_{S^k}P_{Y^{n\star}})$ is equal to the left side of (136). Finally, matching the size of the list to the fidelity of reproduction using (52), we find that L is equal to the right side of (136). ■

If the source is equiprobable, the bound in Theorem 11 becomes particularly simple, as the following result details.

Theorem 12 (Converse, EBMS-BSC)**.** *For $p = \frac{1}{2}$, if there exists a (k,n,d,ϵ) joint source-channel code, then*

$$\lambda\binom{n}{r^\star+1} + \left\langle {n \atop r^\star} \right\rangle \le \left\langle {k \atop \lfloor kd \rfloor} \right\rangle 2^{n-k} \tag{141}$$

where

$$r^\star = \max\left\{ r:\ \sum_{t=0}^{r}\binom{n}{t}\delta^t(1-\delta)^{n-t} \le 1-\epsilon \right\} \tag{142}$$

and $\lambda \in [0,1)$ is the solution to

$$\sum_{j=0}^{r^\star}\binom{n}{t}\delta^t(1-\delta)^{n-t} + \lambda\delta^{r^\star+1}(1-\delta)^{n-r^\star-1}\binom{n}{r^\star+1} = 1-\epsilon \tag{143}$$

The achievability result in Theorem 7 is particularized as follows.

Theorem 13 (Achievability, BMS-BSC)**.** *There exists an* (k, n, d, ϵ) *joint source-channel code with*

$$\epsilon \leq \inf_{M,\gamma>0} \left\{ \mathbb{E}\left[\exp\left\{ - \left| nC - (T_\delta^n - n\delta)\log\frac{1-\delta}{\delta} - \log\frac{\gamma H_M}{\rho(T_p^k)} \right|^+ \right\} \right] \right.$$
$$\left. + e^{-\gamma} + \mathbb{E}\left[(1-\rho(T_p^k))^M \right] \right\} \tag{144}$$

where $\rho \colon \{0, 1, \ldots, k\} \mapsto [0,1]$ *is defined as*

$$\rho(T) = \sum_{t=0}^{k} L(T,t) q^t (1-q)^{n-t} \tag{145}$$

where

$$L(T,t) = \begin{cases} \binom{T}{t_0}\binom{k-T}{t-t_0} & t - kd \leq T \leq t + kd \\ 0 & \textit{otherwise} \end{cases} \tag{146}$$

$$t_0 = \left\lceil \frac{t + T - kd}{2} \right\rceil^+ \tag{147}$$

$$q = \frac{p-d}{1-2d} \tag{148}$$

Proof: We weaken the infima over P_{X^n} and P_{Z^k} in (78) by choosing them to be the product distributions generated by the capacity-achieving channel input distribution and the rate-distortion function-achieving reproduction distribution, respectively, i.e. P_{X^n} is equiprobable on $\{0,1\}^n$, and $P_{Z^k} = P_{\mathsf{Z}^\star} \times \ldots \times P_{\mathsf{Z}^\star}$, where $P_{\mathsf{Z}^\star}(1) = q$. As shown in [12, proof of Theorem 21],

$$P_{Z^k}\left(B_d(s^k)\right) \geq \rho(w(s^k)) \tag{149}$$

On the other hand, $|Y^n - X^n|_0$ is distributed as T_δ^n, so (144) follows by plugging in (135) and (149) into (78). ■

In the special case of the BMS-BSC, Theorem 9 can be strengthened as follows.

Theorem 14 (Gaussian approximation, BMS-BSC)**.** *The parameters of the optimal* (k, n, d, ϵ) *code satisfy* (106) *where* $R(d)$, C, $\mathcal{V}(d)$, V *are given by* (127), (128), (129), (130), *respectively, and the remainder term in* (106) *satisfies*

$$O(1) \leq \theta(n) \tag{150}$$
$$\leq \left(1 + \frac{1}{\log e}\right)\log k + \log\log k + O(1) \tag{151}$$

if $0 < d < p$*, and*

$$-\frac{1}{2}\log n + O\left(1\right) \leq \theta\left(n\right) \tag{152}$$

$$\leq \frac{1}{2}\log n + O\left(1\right) \tag{153}$$

if $d = 0$.

Proof: An asymptotic analysis of the converse bound in Theorem 11 akin to that found in [12, proof of Theorem 23] leads to (150) and (152). An asymptotic analysis of the achievability bound in Theorem 13 similar to the one found in [12, Appendix G] leads to (151). Finally, (153) is the same as (117). ■

The bounds and the Gaussian approximation (in which we take $\theta\left(n\right) = 0$ are plotted in Fig. 3 ($d = 0$), Fig. 4 (fair binary source, $d > 0$) and Fig. 5 (biased binary source, $d > 0$). A source of fair coin flips has zero dispersion, and as anticipated in Remark 5, JSSC does not afford much gain in the finite blocklength regime (Fig. 4). The situation is different if the source is biased, with JSCC showing significant gain over SSCC (Figures 3 and 5).

VII. Transmission of a GMS over an AWGN channel

In this section we analyze the setup where the Gaussian memoryless source $S_i \sim \mathcal{N}(0, \sigma_{\mathsf{S}}^2)$ is transmitted over an AWGN channel, which, upon receiving an input x^n, outputs $Y^n = x^n + N^n$, where $N^n \sim \mathcal{N}(0, \sigma_{\mathsf{N}}^2\mathbf{I})$. The encoder/decoder must satisfy two constraints, the fidelity constraint and the cost constraint:

- the MSE distortion exceeds $0 \leq d \leq \sigma_{\mathsf{S}}^2$ with probability no greater than $0 < \epsilon < 1$;
- each channel codeword satisfies the equal power constraint in (9).[5]

The capacity-cost function and the rate-distortion function are given by

$$R(d) = \frac{1}{2}\log\left(\frac{\sigma_{\mathsf{S}}^2}{d}\right) \tag{154}$$

$$C(P) = \frac{1}{2}\log\left(1 + P\right) \tag{155}$$

[5]See Remark 9 in Section V for a discussion of the close relation between an equal and a maximal power constraint.

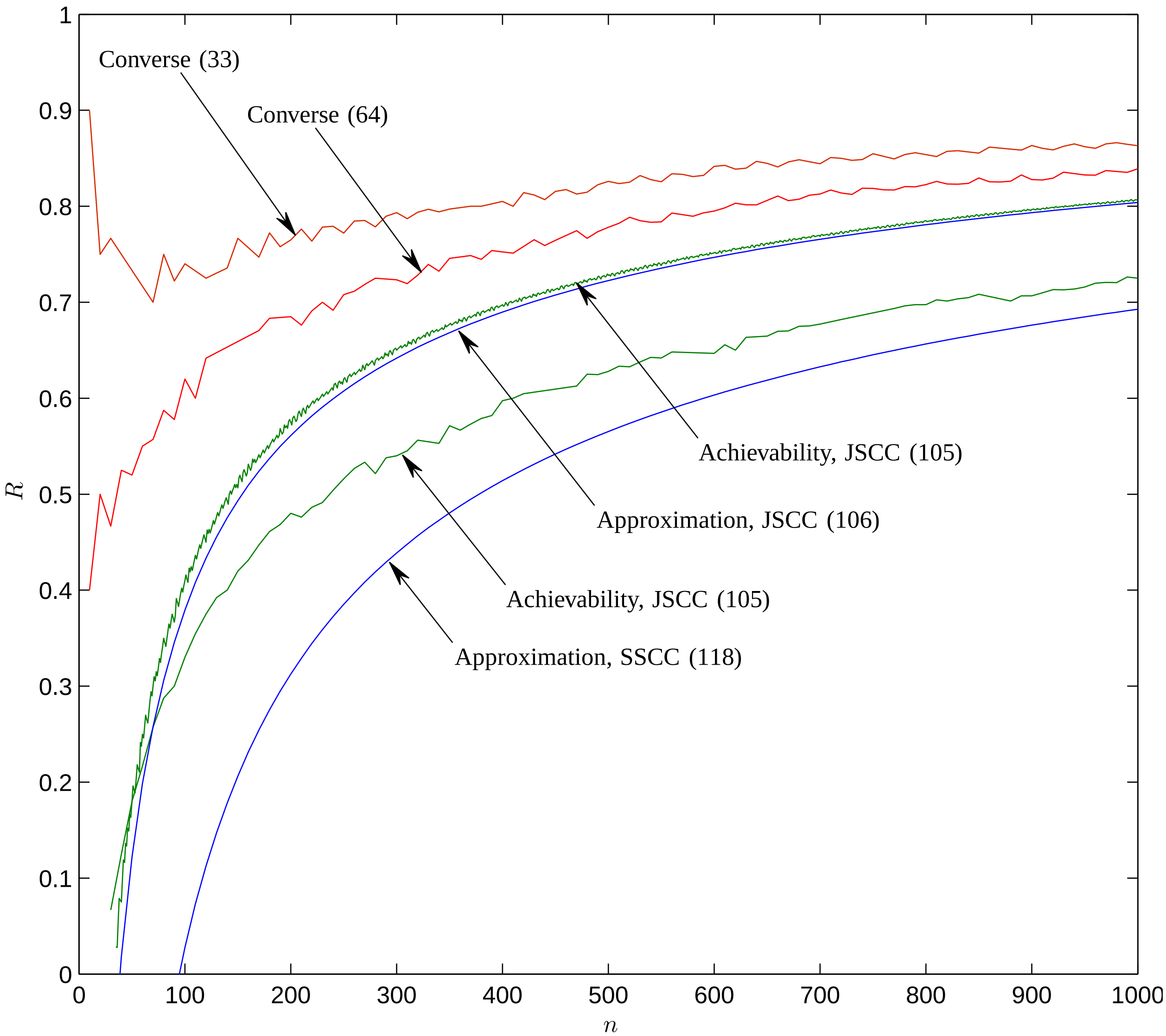


Fig. 3. Rate-blocklength tradeoff for the transmission of a BMS with bias $p = 0.11$ over a BSC with crossover probability $\delta = p = 0.11$ and $d = 0$, $\epsilon = 10^{-2}$.

The source dispersion is given by [12]:

$$\mathcal{V}(d) = \frac{1}{2} \log^2 e \tag{156}$$

while the channel dispersion is given by (110) [14].

In the rest of the section, W_λ^ℓ denotes a noncentral chi-square distributed random variable with ℓ degrees of freedom and non-centrality parameter λ, independent of all other random variables. The Euclidean norm is denoted by $|\cdot|$.

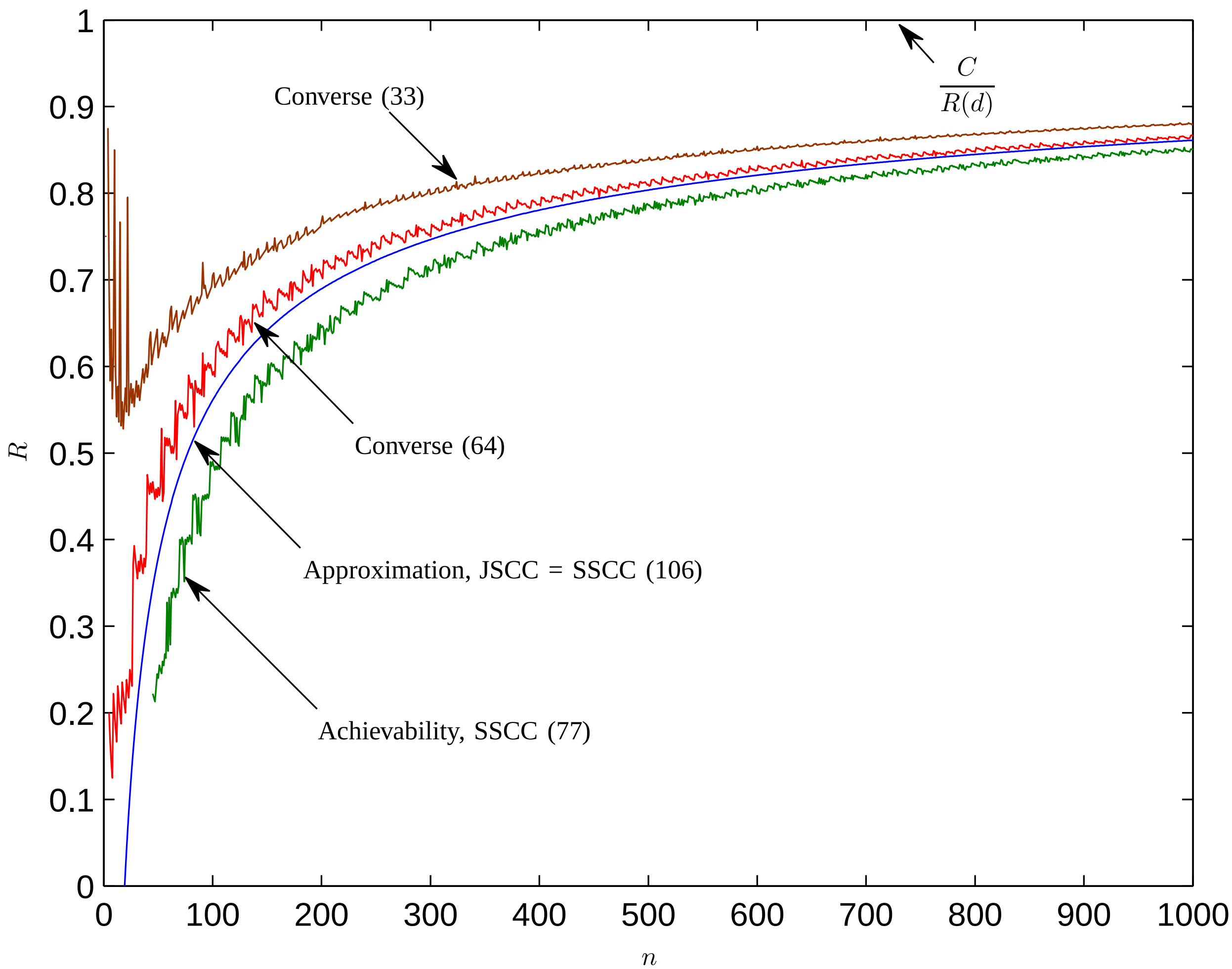


Fig. 4. Rate-blocklength tradeoff for the transmission of a fair BMS over a BSC with crossover probability $\delta = d = 0.11$ and $\epsilon = 10^{-2}$.

A straightforward particularization of the d-tilted information converse in Theorem 2 leads to the following result.

Theorem 15 (Converse, GMS-AWGN)**.** *If there exists a* (k, n, d, ϵ) *code, then*

$$\epsilon \geq \sup_{\gamma \geq 0} \left\{ \mathbb{P}\left[\frac{\log e}{2} \left(W_0^k - k \right) + \frac{\log e}{2} \left(\frac{P}{1+P} W_{\frac{n}{P}}^n - n \right) \geq nC(P) - kR(d) + \gamma \right] - \exp\left(-\gamma\right) \right\} \tag{157}$$

Observe that the terms to the left of the '$\geq$' sign inside the probability in (157) are zero-mean random variables whose variances are equal to $k\mathcal{V}(d)$ and nV, respectively.

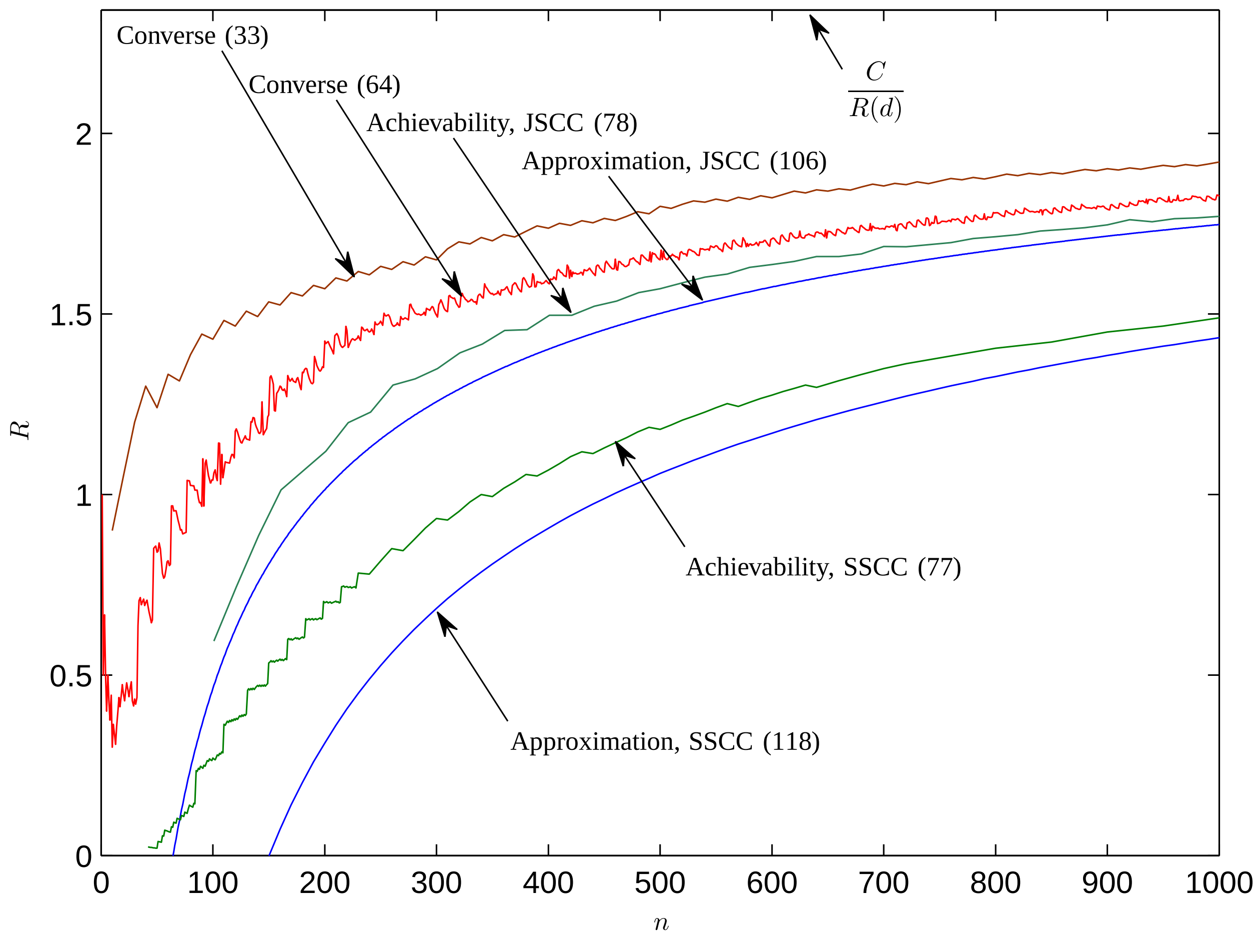


Fig. 5. Rate-blocklength tradeoff for the transmission of a BMS with bias $p = 0.11$ over a BSC with crossover probability $\delta = p = 0.11$ and $d = 0.05$, $\epsilon = 10^{-2}$.

Proof: The spherically-symmetric $P_{\bar{Y}^n} = P_{Y^{n\star}} = P_{\mathsf{Y}^\star} \times \ldots \times P_{\mathsf{Y}^\star}$, where $\mathsf{Y}^\star \sim \mathcal{N}(0, \sigma_{\mathsf{N}}^2(1+P))$ is the capacity-achieving output distribution, satisfies the symmetry assumption of Theorem 2. More precisely, it is not hard to show (see [14, (205)]) that for all $x^n \in \mathcal{F}(\alpha)$, $\imath_{X^n;Y^{n\star}}(x^n; Y^n)$ has the same distribution under $P_{Y^{n\star}|X^n=x^n}$ as

$$\frac{n}{2}\log(1+P) - \frac{\log e}{2}\left(\frac{P}{1+P}W_{\frac{n}{P}}^{n} - n\right) \tag{158}$$

The d-tilted information in s^k is given by

$$\jmath_{S^k}(s^k, d) = \frac{k}{2}\log\frac{\sigma_{\mathsf{S}}^2}{d} + \left(\frac{|s^k|^2}{\sigma_{\mathsf{S}}^2} - k\right)\frac{\log e}{2} \tag{159}$$

Plugging (158) and (159) into (33), (157) follows. ∎

The hypothesis testing converse in Theorem 5 is particularized as follows.

Theorem 16 (HT converse, GMS-AWGN)**.**

$$k\int_0^\infty r^{k-1}\mathbb{P}\left[PW^n_{n\left(1+\frac{1}{P}\right)}+k\frac{d}{\sigma^2}r^2\le n\tau\right]dr\le 1 \tag{160}$$

where τ is the solution to

$$\mathbb{P}\left[\frac{P}{1+P}W^n_{\frac{n}{P}}+W^k_0\le n\tau\right]=1-\epsilon \tag{161}$$

Proof: As in the proof of Theorem 15, we let $\bar{Y}^n\sim Y^{n\star}\sim\mathcal{N}(0,\sigma^2_{\mathsf{N}}(1+P)\mathbf{I})$. Under $P_{Y^n|X^n=x^n}$, the distribution of $\imath_{X^n;Y^{n\star}}(x^n;Y^{n\star})$ is that of (158), while under $P_{Y^{n\star}}$, it has the same distribution as (cf. [14, (204)])

$$\frac{n}{2}\log(1+P)-\frac{\log e}{2}\left(PW^n_{n\left(1+\frac{1}{P}\right)}-n\right) \tag{162}$$

Since the distribution of $\imath_{X^n;Y^{n\star}}(x^n;Y^{n\star})$ does not depend on the choice of $x^n\in\mathbb{R}^n$ according to either measure, Theorem 5 applies. Further, choosing Q_{S^k} to be the Lebesgue measure on $\mathbb{R}^k$, i.e. $dQ_{S^k}=ds^k$, observe that

$$\log f_{S^k}(s^k)=\log\frac{dP_{S^k}(s^k)}{ds^k}=-\frac{k}{2}\log\left(2\pi\sigma^2_{\mathsf{S}}\right)-\frac{\log e}{2\sigma^2_{\mathsf{S}}}|s^k|^2 \tag{163}$$

Now, (160) and (161) are obtained by integrating

$$1\left\{\log f_{S^k}(s^k)+\imath_{X^n;Y^{n\star}}(x^n;y^n)>\ \frac{n}{2}\log(1+P)+\frac{n}{2}\log e-\frac{k}{2}\log(2\pi\sigma^2_{\mathsf{S}})-\frac{\log e}{2}n\tau\right\} \tag{164}$$

with respect to $ds^kdP_{Y^{n\star}}(y^n)$ and $dP_{S^k}(s^k)dP_{Y^n|X^n=x^n}(y^n)$, respectively. ■

The bound in Theorem 7 can be computed as follows.

Theorem 17 (Achievability, GMS-AWGN)**.** *There exists a (k,n,d,ϵ) code such that*

$$\begin{aligned}\epsilon\ \le\inf_{M,\gamma>0}&\left\{\mathbb{E}\left[\exp\left\{-\left|nC(P)-\frac{\log e}{2}\left(\frac{P}{1+P}W^n_{\frac{n}{P}}-n\right)-\log\frac{\gamma FH_M}{\rho(W^k_0)}\right|^+\right\}\right]\right.\\&\left.+e^{-\gamma}+\mathbb{E}\left[\left(1-\rho\left(W^k_0\right)\right)^M\right]\right\}\end{aligned} \tag{165}$$

where

$$F=\max_{n\in\mathbb{N},t\in\mathbb{R}^+}\frac{f_{W^n_{nP}}(t)}{f_{W^n_0}\left(\frac{t}{1+P}\right)}<\infty \tag{166}$$

f_W is the probability density function of W, and $\rho : \mathbb{R}^+ \mapsto [0,1]$ is defined by

$$\rho(t) = \frac{\Gamma\left(\frac{k}{2}+1\right)}{\sqrt{\pi}k\Gamma\left(\frac{k-1}{2}+1\right)} \left(1 - L\left(\sqrt{\frac{t}{k}}\right)\right)^{\frac{k-1}{2}} \tag{167}$$

where

$$L(r) = \begin{cases} 0 & r < \sqrt{\frac{d}{\sigma_{\mathsf{S}}^2}} - \sqrt{1-\frac{d}{\sigma_{\mathsf{S}}^2}} \\ 1 & \left|r - \sqrt{1-\frac{d}{\sigma_{\mathsf{S}}^2}}\right| > \sqrt{\frac{d}{\sigma_{\mathsf{S}}^2}} \\ \frac{\left(1+r^2-2\frac{d}{\sigma_{\mathsf{S}}^2}\right)^2}{4\left(1-\frac{d}{\sigma_{\mathsf{S}}^2}\right)r^2} & \textit{otherwise} \end{cases} \tag{168}$$

Proof: We compute an upper bound to (78) for the specific case of the GMS over an AWGN channel. First, we weaken the infimum over P_{Z^k} in (78) by choosing P_{Z^k} to be the uniform distribution on the surface of the k-dimensional sphere with center at $\mathbf{0}$ and radius $r_0 = \sqrt{k}\sigma\sqrt{1-\frac{d}{\sigma_{\mathsf{S}}^2}}$. It was shown in [12, proof of Theorem 37] (see also [6], [18]),

$$P_{Z^k}\left(B_d(s^k)\right) \geq \rho\left(|s^k|^2\right) \tag{169}$$

which takes care of the source random variable in (78).

Now let us consider the channel random variable $\imath_{X^n;Y^n}(X^n;Y^n)$. Observe that since X^n lies on the power sphere and the noise is spherically symmetric, $|Y^n|^2 = |X^n+N^n|^2$ has the same distribution as $|x_0^n+N^n|^2$, where x_0^n is an arbitrary point on the surface of the power sphere. Letting $x_0^n = \sigma_{\mathsf{N}}\sqrt{P}\,(1,1,\ldots,1)$, we see that $\frac{1}{\sigma_{\mathsf{N}}}|x_0^n+N^n|^2 = \sum_{i=1}^n\left(\frac{1}{\sigma_{\mathsf{N}}^2}N_i + \sqrt{P}\right)^2$ has noncentral chi-squared distribution with n degrees of freedom and noncentrality parameter nP. To simplify calculations, we express the information density as

$$\imath_{X^n;Y^n}(x_0^n;y^n) = \imath_{X^n;Y^{n\star}}(x_0^n;y^n) - \frac{dP_{Y^n}}{dP_{Y^{n\star}}}\left(y^n\right) \tag{170}$$

where $Y^{n\star} \sim \mathcal{N}(0, \sigma_{\mathsf{N}}^2(1+P)\mathbf{I})$. The distribution of $\imath_{X^n;Y^{n\star}}(x_0^n;Y^n)$ is the same as (158). Further, due to the spherical symmetry of both P_{Y^n} and $P_{Y^{n\star}}$, as discussed above, we have

$$\frac{dP_{Y^n}}{dP_{Y^{n\star}}}\left(Y^n\right) \sim \frac{f_{W_{nP}^n}\left(W_{nP}^n\right)}{f_{W_0^n}\left(\frac{W_{nP}^n}{1+P}\right)} \tag{171}$$

which is bounded uniformly in n as observed in [14, (425), (435)], thus (166) is finite, and (165) follows. ∎

The following result strengthens Theorem 9 in the special case of the GMS-AWGN.

Theorem 18 (Gaussian approximation, GMS-AWGN)**.** *The parameters of the optimal* (k, n, d, ϵ) *code satisfy* (106) *where* $R(d)$, C, $\mathcal{V}(d)$, V *are given by* (154), (155), (156), (110), *respectively, and the remainder term in* (106) *satisfies*

$$O(1) \leq \theta(n) \tag{172}$$

$$\leq \left(1 + \frac{1}{\log e}\right) \log k + \log\log k + O(1) \tag{173}$$

Proof: An asymptotic analysis of the converse bound in Theorem 16 similar to that found in [12, proof of Theorem 40] leads to (172). An asymptotic analysis of the achievability bound in Theorem 17 similar to [12, Appendix K] leads to (173). ∎

While their numerical evaluation reveals that bounds are not as tight as those in Section VI, they suffice to show that JSCC noticeably outperforms SSCC in the displayed region of blocklengths (Fig. 6).

VIII. To code or not to code

In this section, we compare the performance of blocklength-1 codes with that of the best blocklength-n codes, leveraging the bounds in Sections III and IV and the approximation in Section V. We show certain examples when symbol-by-symbol coding is, in fact, either optimal or very close to being optimal.

A. Performance of symbol-by-symbol source-channel codes

Definition 9. *An* (n, d, ϵ, α) *symbol-by-symbol code is an* $(n, n, d, \epsilon, \alpha)$ *code* (f, g) *(according to Definition 1) that satisfies*

$$\mathsf{f}(s^n) = (\mathsf{f}_1(s_1), \ldots, \mathsf{f}_1(s_n)) \tag{174}$$

$$\mathsf{g}(y^n) = (\mathsf{g}_1(y_1), \ldots, \mathsf{g}_1(y_n)) \tag{175}$$

for some pair of functions $\mathsf{f}_1 \colon \mathcal{S} \mapsto \mathcal{A}$ *and* $\mathsf{g}_1 \colon \mathcal{B} \mapsto \hat{\mathcal{S}}$.

The minimum excess distortion achievable with symbol-by-symbol codes at channel blocklength n, *excess probability* ϵ *and cost* α *is defined by*

$$D_1(n, \epsilon, \alpha) = \inf\left\{d \colon \exists (n, d, \epsilon, \alpha) \text{ symbol-by-symbol code}\right\}. \tag{176}$$

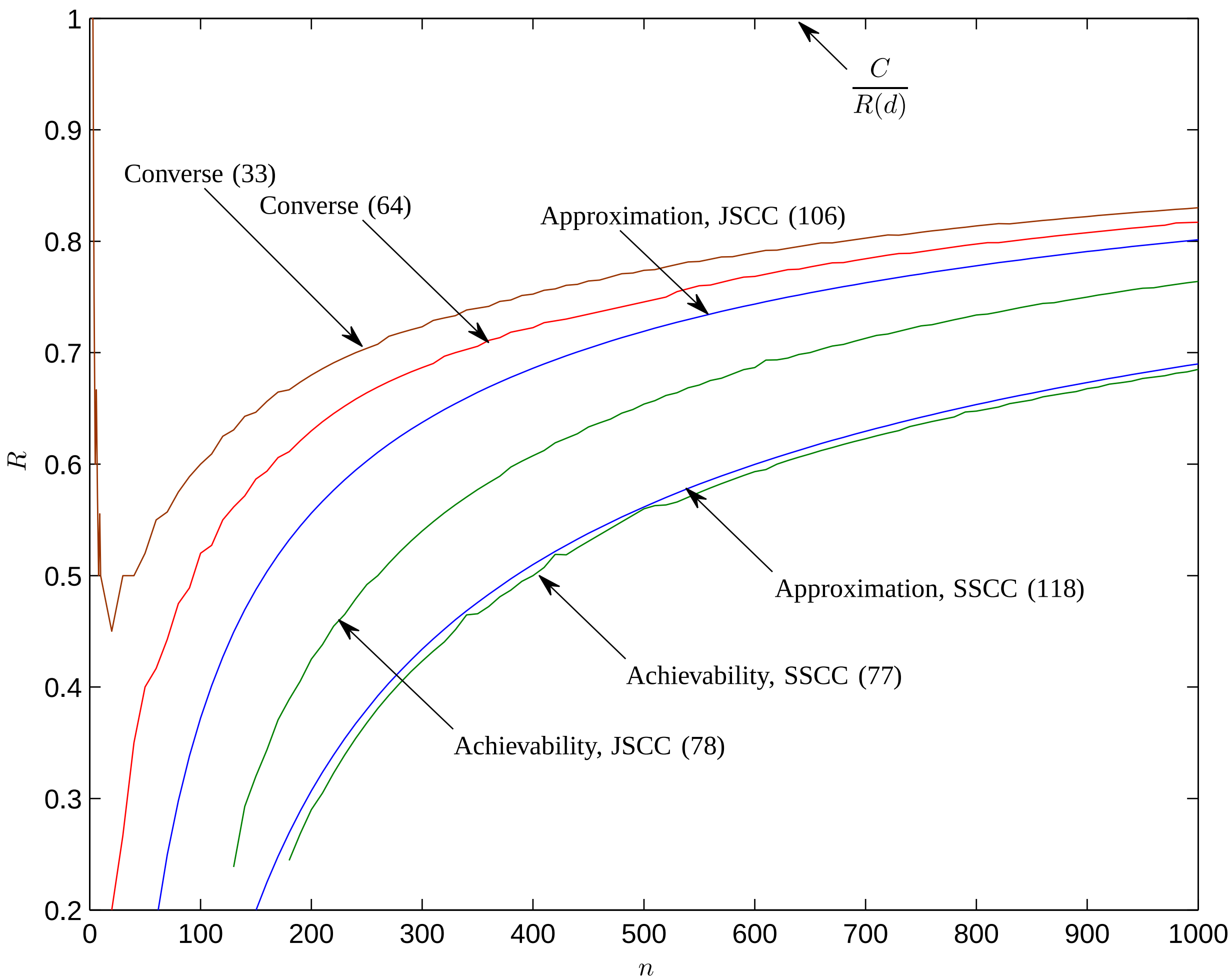


Fig. 6. Rate-blocklength tradeoff for the transmission of a GMS with $\frac{d}{\sigma_s^2} = 0.5$ over an AWGN channel with $P = 1$, $\epsilon = 10^{-2}$.

Definition 10. *The distortion-dispersion function of symbol-by-symbol joint source-channel coding is defined as*

$$\mathcal{W}_1(\alpha) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} n \left(\frac{D\left(C(\alpha)\right) - D_1(n, \epsilon, \alpha)}{Q^{-1}\left(\epsilon\right)} \right)^2 \tag{177}$$

where $D(\cdot)$ is the distortion-rate function of the source.

As before, if there is no cost constraint ($\mathsf{c}^n(x^n) = 0$ for all $x \in \mathcal{A}^n$), we will simplify the notation and write $D_1(n, \epsilon)$ for $D_1(n, \epsilon, \alpha)$ and $\mathcal{W}_1$ for $\mathcal{W}_1(\alpha)$.

A symbol-by-symbol code has rate 1. Our goal in this section is to compare the excess

distortion performance of the optimal code of rate 1 at channel blocklength n with that of the optimal symbol-by-symbol code, evaluated after n channel uses.

In addition to restrictions (i)–(iv) of Section V, we assume that the channel and the source are probabilistically matched in the following sense (cf. [11]).

(v) There exist α, $P_{\mathsf{X}^\star|\mathsf{S}}$ and $P_{\mathsf{Z}^\star|\mathsf{Y}}$ such that $P_{\mathsf{X}^\star}$ and $P_{\mathsf{Z}^\star|\mathsf{S}}$ generated by the joint distribution $P_{\mathsf{S}}P_{\mathsf{X}^\star|\mathsf{S}}P_{\mathsf{Y}|\mathsf{X}}P_{\mathsf{Z}^\star|\mathsf{Y}}$ achieve the capacity-cost function $C(\alpha)$ and the distortion-rate function $D\left(C(\alpha)\right)$, respectively. The channel cost function is assumed to be separable, $\mathsf{c}_n(x^n) = \frac{1}{n}\sum_{i=1}^n \mathsf{c}(x_i)$.

Condition (v) ensures that symbol-by-symbol transmission attains the minimum average (over source realizations) distortion achievable among all codes of any blocklength. The following results pertain to the full distribution of the distortion incurred at the receiver output and not just its mean.

Theorem 19 (Achievability, symbol-by-symbol code)**.** *Under restrictions* (i)-(v)*, if*

$$\mathbb{P}\left[\sum_{i=1}^n \mathsf{d}(S_i, Z_i^\star) > nd\right] \leq \epsilon \tag{178}$$

where $P_{Z^{n\star}|S^n} = P_{\mathsf{Z}^\star|\mathsf{S}} \times \ldots \times P_{\mathsf{Z}^\star|\mathsf{S}}$*, and* $P_{\mathsf{Z}^\star|\mathsf{S}}$ *achieves* $D\left(C(\alpha)\right)$*, then there exists an* (n, d, ϵ, α) *symbol-by-symbol code (average cost constraint).*

Proof: As shown in [11], if (v) holds there exist a symbol-by-symbol encoder and decoder such that the conditional distribution of the output of the decoder given the source outcome coincides with distribution $P_{\mathsf{Z}^\star|\mathsf{S}}$, so the excess-distortion probability of this symbol-by-symbol code is given by (178). ■

Theorem 20 (Converse, symbol-by-symbol code)**.** *Under restriction* (i)*, any* (n, d, ϵ, α) *symbol-by-symbol code (average cost constraint) must satisfy*

$$\epsilon \geq \inf_{\substack{P_{\mathsf{Z}|\mathsf{S}}:\\ I(\mathsf{S};\mathsf{Z}) \leq C(\alpha)}} \mathbb{P}\left[\mathsf{d}^n(S^n, Z^n) > d\right] \tag{179}$$

where $P_{Z^n|S^n} = P_{\mathsf{Z}|\mathsf{S}} \times \ldots \times P_{\mathsf{Z}|\mathsf{S}}$.

Proof: The excess-distortion probability at blocklength n, distortion d and cost α achievable

among all single-letter codes $P_{\mathsf{X}|\mathsf{S}}$, $P_{\mathsf{Z}|\mathsf{Y}}$ must satisfy

$$\epsilon \geq \inf_{\substack{P_{\mathsf{X}|\mathsf{S}}, P_{\mathsf{Z}|\mathsf{Y}}:\\ \mathsf{S}-\mathsf{X}-\mathsf{Y}-\mathsf{Z}\\ \mathbb{E}[\mathsf{c}(\mathsf{X})]\leq\alpha}} \mathbb{P}\left[\mathsf{d}^n(S^n, Z^n) > d\right] \tag{180}$$

$$\geq \inf_{\substack{P_{\mathsf{X}|\mathsf{S}}, P_{\mathsf{Z}|\mathsf{Y}}:\\ \mathbb{E}[\mathsf{c}(\mathsf{X})]\leq\alpha\\ I(\mathsf{S};\mathsf{Z})\leq I(\mathsf{X};\mathsf{Y})}} \mathbb{P}\left[\mathsf{d}^n(S^n, Z^n) > d\right] \tag{181}$$

where (181) holds since $\mathsf{S}-\mathsf{X}-\mathsf{Y}-\mathsf{Z}$ implies $I(\mathsf{S};\mathsf{Z}) \leq I(\mathsf{X};\mathsf{Y})$ by the data processing inequality. The right side of (181) is lower bounded by the right side of (179) because $I(\mathsf{X};\mathsf{Y}) \leq C(\alpha)$ holds for all P_{X} with $\mathbb{E}\left[\mathsf{c}(\mathsf{X})\right] \leq \alpha$. ■

Theorem 21 (Gaussian approximation, optimal symbol-by-symbol code)**.** *Assume* $\mathbb{E}\left[\mathsf{d}^3\left(\mathsf{S},\mathsf{Z}\right)\right] < \infty$*. Under restrictions* (i)-(v)*,*

$$D_1(n,\epsilon,\alpha) = D\left(C(\alpha)\right) + \sqrt{\frac{\mathcal{W}_1(\alpha)}{n}}Q^{-1}\left(\epsilon\right) + \frac{\theta_1(n)}{n} \tag{182}$$

$$\mathcal{W}_1(\alpha) = \mathrm{Var}\left[\mathsf{d}(\mathsf{S},\mathsf{Z}^\star)\right] \tag{183}$$

where

$$\theta_1(n) \leq O\left(1\right) \tag{184}$$

Moreover, if there is no power constraint,

$$\theta_1(n) \geq \frac{D'(R)}{R^2}\theta(n) \tag{185}$$

$$\mathcal{W}_1 = \mathcal{W}(1) \tag{186}$$

where $\theta(n)$ is that in Theorem 9.

If $\mathrm{Var}\left[\mathsf{d}\left(\mathsf{S},\mathsf{Z}\right)\right] > 0$ *and $\mathcal{S}$, $\hat{\mathcal{S}}$ are finite, then*

$$\theta_1(n) \geq O\left(1\right) \tag{187}$$

Proof: Since the third absolute moment of $\mathsf{d}(S_i, Z_i^\star)$ is finite, the achievability part of (182), namely, (182) with the remainder satisfying (184), follows by a straightforward application of the Berry-Esseen bound to (178), provided that $\mathrm{Var}\left[\mathsf{d}(S_i, Z_i^\star)\right] > 0$. If $\mathrm{Var}\left[\mathsf{d}(S_i, Z_i^\star)\right] = 0$, it follows trivially from (178).

To show the converse in (185), observe that since the set of all (n, n, d, ϵ) codes includes all (n, d, ϵ) symbol-by-symbol codes, we have $D(n, n, \epsilon) \leq D_1(n, \epsilon)$. Since $Q^{-1}(\epsilon)$ is positive or negative depending on whether $\epsilon < \frac{1}{2}$ or $\epsilon > \frac{1}{2}$, using (122) we conclude that we must necessarily have (186), which is, in fact, a consequence of conditions (b), (c) in Section II and (v). Now, (185) is simply the converse part of (121).

The proof of the refined converse in (187) is relegated to Appendix E. ■

Theorem 21 shows that if the source and the channel are probabilistically matched in the sense of [11], then not only does symbol-by-symbol transmission achieve the minimum average distortion, but also the dispersion of JSCC. In other words, not only do such symbol-by-symbol codes attain the minimum average distortion but also the variance of distortions at the decoder's output is the minimum achievable among all codes operating at that average distortion.

Two conspicuous examples that satisfy the probabilistic matching condition (v), so that symbol-by-symbol coding is optimal in terms of average distortion, are the transmission of a binary equiprobable source over a binary-symmetric channel provided the desired bit error rate is equal to the crossover probability of the channel [19, Sec.11.8], [20, Problem 7.16], and the transmission of a Gaussian source over an additive white Gaussian noise channel under the mean-square error distortion criterion, provided that the tolerable source signal-to-noise ratio attainable by an estimator is equal to the signal-to-noise ratio at the output of the channel [21]. We dissect these two examples next.

B. Symbol-by-symbol coding for lossy transmission of BMS over BSC

In the setup of Section VI, if the source is unbiased $\left(p = \frac{1}{2}\right)$, then $C = 1 - h(\delta)$, $R(d) = 1 - h(d)$, and $D(C) = \delta$. If the encoder and the decoder are both identity mappings (uncoded transmission), the resulting joint distribution satisfies condition (v). Using (122) and (183), it is easy to verify that

$$\mathcal{W}(1) = \mathcal{W}_1 = \delta(1 - \delta) \tag{188}$$

that is, uncoded transmission is optimal in terms of dispersion, as anticipated in (186).

Moreover, regardless of the allowed ϵ, uncoded transmission attains the minimum distortion $D(n, n, \epsilon)$ achievable among all codes operating at blocklength n, as the following result demonstrates.

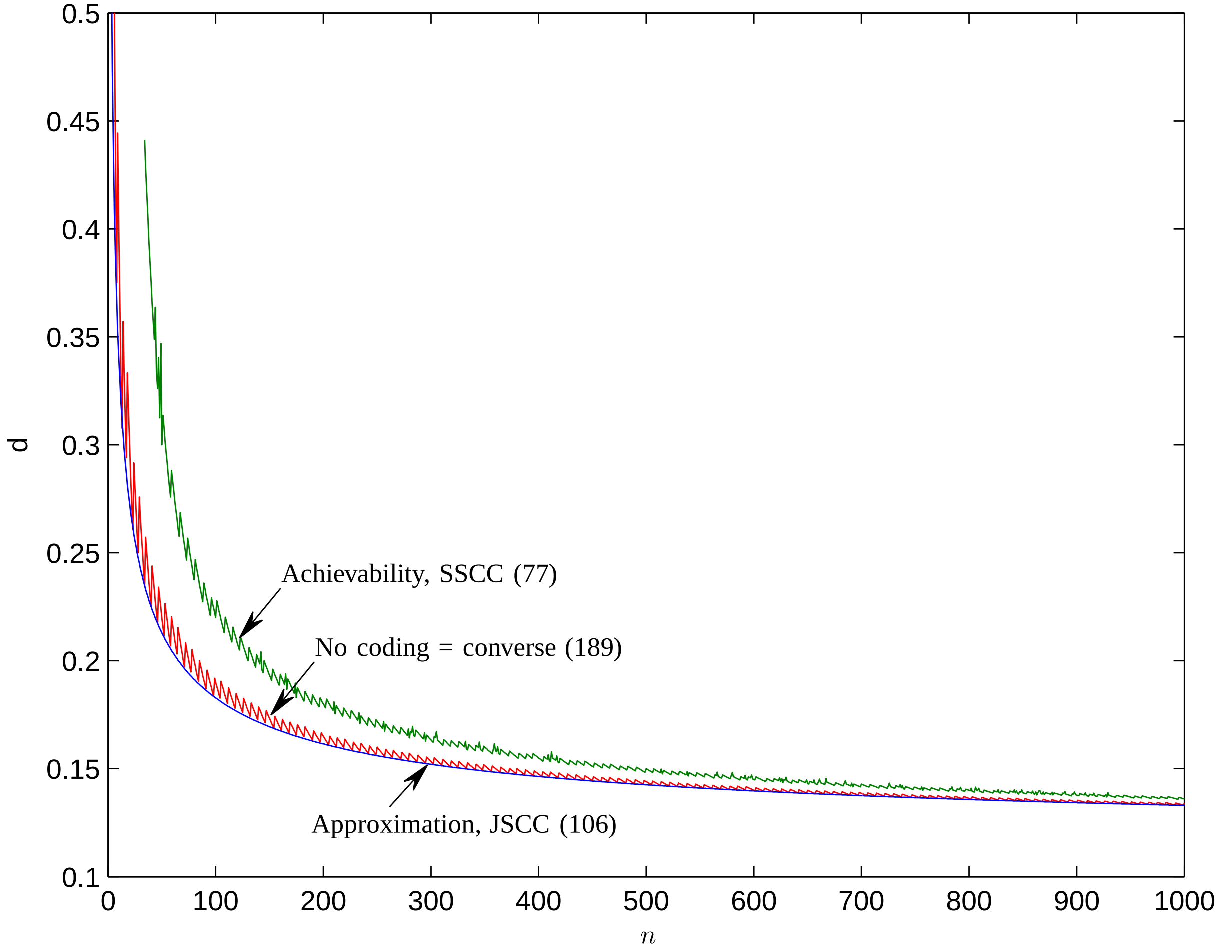


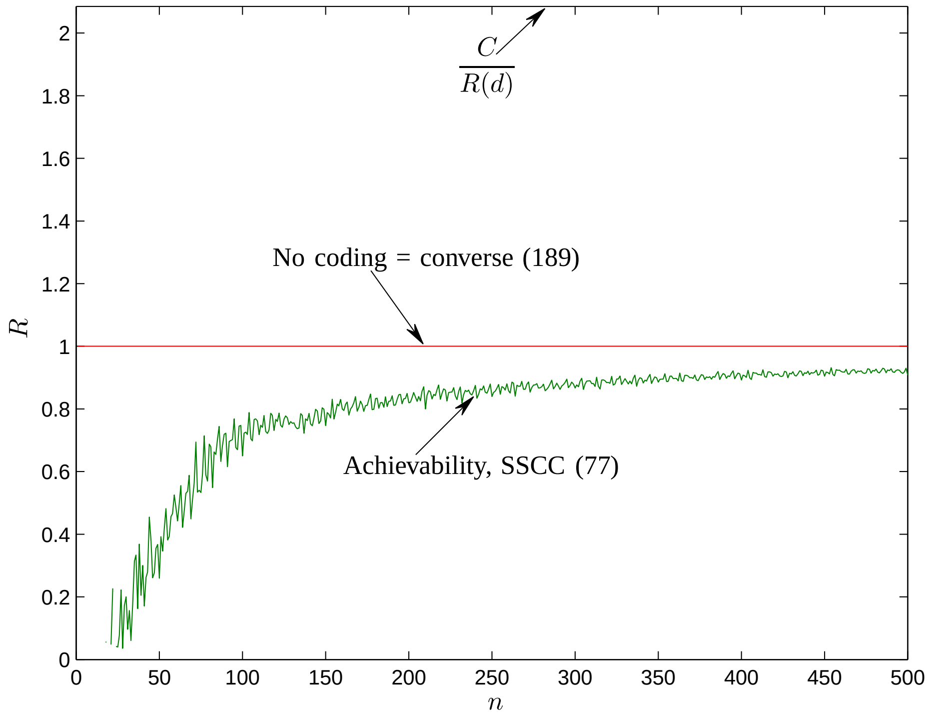
Fig. 7. Distortion-blocklength tradeoff for the transmission of a fair BMS over a BSC with crossover probability $\delta = 0.11$ and $R = 1$, $\epsilon = 10^{-2}$.

Theorem 22 (BMS-BSC, symbol-by-symbol code)**.** *At blocklength n and excess distortion probability ϵ, the uncoded scheme achieves, regardless of $0 \leq p \leq 1$, $\delta \leq \frac{1}{2}$,*

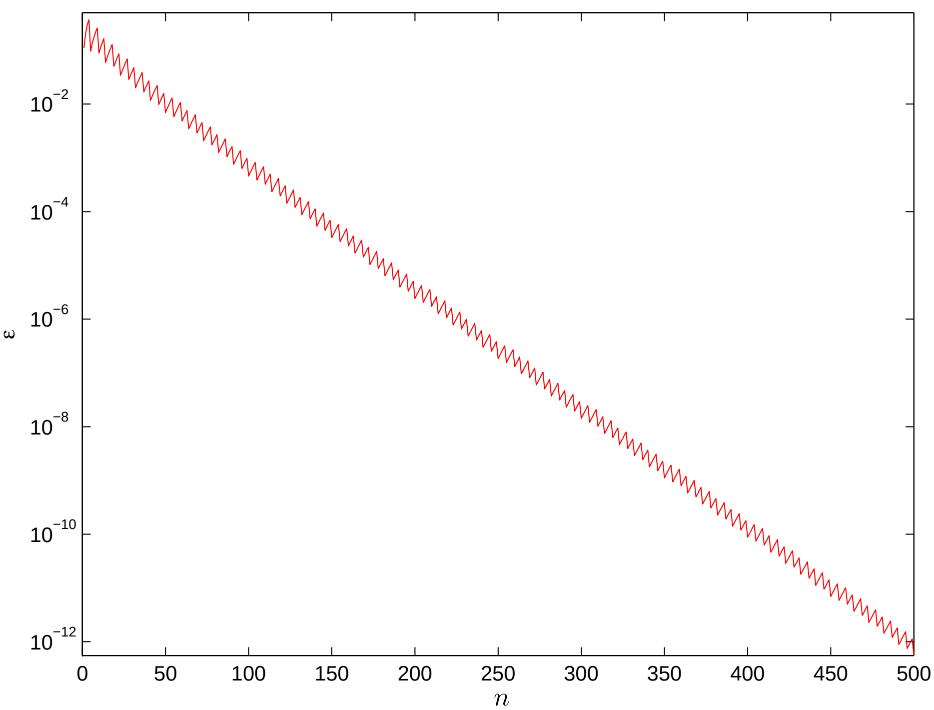
$$D_1(n, \epsilon) = \min \left\{ d : \sum_{t=0}^{\lfloor nd \rfloor} \binom{n}{t} \min\{p, \delta\}^t (1 - \min\{p, \delta\})^{n-t} \geq 1 - \epsilon \right\} \tag{189}$$

Moreover, if the source is equiprobable $\left(p = \frac{1}{2}\right)$,

$$D_1(n, \epsilon) = D(n, n, \epsilon) \tag{190}$$



(a)



(b)

Fig. 8. Rate-blocklength tradeoff (a) for the transmission of a fair BMS over a BSC with crossover probability $\delta = 0.11$ and $d = 0.22$. The excess-distortion probability ϵ is set to be the one achieved by the uncoded scheme (b).

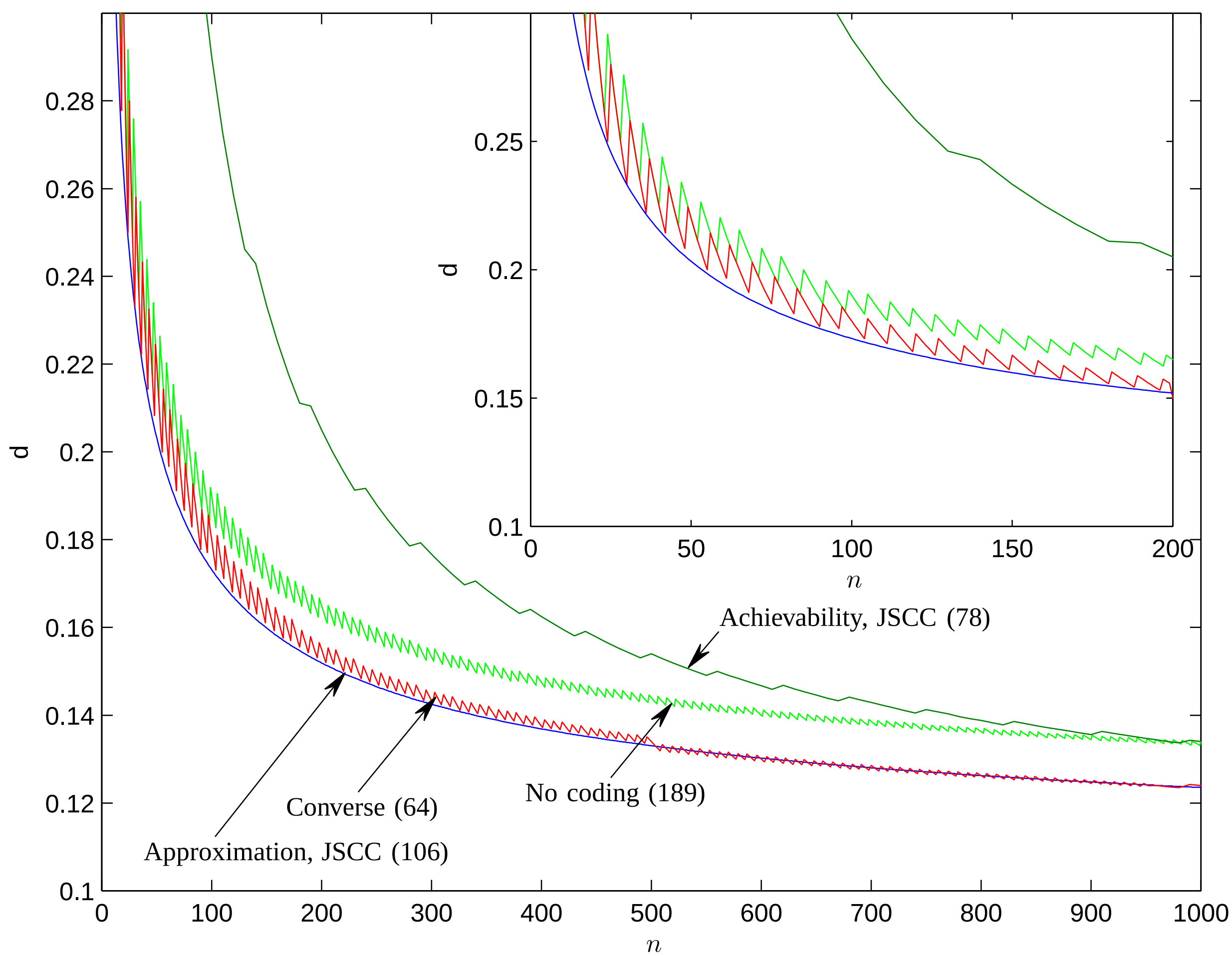


Fig. 9. Distortion-blocklength tradeoff for the transmission of a BMS with $p = \frac{2}{5}$ over a BSC with crossover probability $\delta = 0.11$ and $R = 1$, $\epsilon = 10^{-2}$.

Proof: Direct calculation yields (189). To show (190), let us compare $d^\star = D_1(n, \epsilon)$ with the conditions imposed on d by Theorem 12. Comparing (189) to (142), we see that either

(a) equality in (189) is achieved, $r^\star = nd^\star$, $\lambda = 0$, and (plugging $k = n$ into (141))

$$\left\langle {n \atop nd^\star} \right\rangle \leq \left\langle {n \atop \lfloor nd \rfloor} \right\rangle \tag{191}$$

thereby implying that $d \geq d^\star$, or

(b) $r^\star = nd^\star - 1$, $\lambda > 0$, and (141) becomes

$$\lambda \binom{n}{nd^\star} + \left\langle {n \atop nd^\star - 1} \right\rangle \leq \left\langle {n \atop \lfloor nd \rfloor} \right\rangle \tag{192}$$

which also implies $d \geq d^\star$. To see this, note that $d < d^\star$ would imply $\lfloor nd \rfloor \leq nd^\star - 1$ since

$nd^\star$ is an integer, which in turn would require (according to (192)) that $\lambda \leq 0$, which is impossible.

■

For the transmission of the fair binary source over a BSC, Figure 7 shows the distortion achieved by the uncoded scheme and the separated scheme versus n for a fixed excess-distortion probability $\epsilon = 0.01$. Figure 8(a) shows the rate achieved by separate coding when $d > \delta$ is fixed, and the excess-distortion probability ϵ, shown in Fig. 8(b), is set to be the one achieved by uncoded transmission, namely, (189). Figure 8(a) highlights the fact that at short blocklengths (say $n \leq 100$) separate source/channel coding is vastly suboptimal. As the blocklength increases, the performance of the separated scheme approaches that of the no-coding scheme, but according to Theorem 22 it can never outperform it. Had we allowed the excess distortion probability to vanish sufficiently slowly, the JSCC curve would have approached the Shannon limit as $n \to \infty$. However, in Figure 8(a), the exponential decay in ϵ is such that there is indeed an asymptotic rate penalty as predicted in [3].

For the biased binary source with $p = \frac{2}{5}$ and BSC with crossover probability 0.11, Figure 9 plots the maximum distortion achieved with probability 0.99 by the uncoded scheme, which in this case is asymptotically suboptimal. Nevertheless, uncoded transmission performs remarkably well in the displayed range of blocklengths, achieving the converse almost exactly at blocklengths less than 100, and outperforming the JSCC achievability result in Theorem 13 at blocklengths as long as 900. This example substantiates that even in the absence of a probabilistic match between the source and the channel, symbol-by-symbol transmission, though asymptotically suboptimal, might outperform SSCC and even our random JSCC achievability bound in the finite blocklength regime.

C. Symbol-by-symbol coding for lossy transmission of a GMS over an AWGN

In the setup of Section VII, using (154) and (155), we find that

$$D(C(P)) = \frac{\sigma_{\mathsf{S}}^2}{1+P} \tag{193}$$

The next result characterizes the distribution of the distortion incurred by the symbol-by-symbol scheme that attains the minimum average distortion.

Theorem 23 (GMS-AWGN, symbol-by-symbol code)**.** *The following symbol-by-symbol transmission scheme in which the encoder and the decoder are the amplifiers:*

$$\mathsf{f}_1(\mathsf{s}) = a\mathsf{s},\ \ a^2 = \frac{P\sigma_{\mathsf{N}}^2}{\sigma_{\mathsf{S}}^2} \tag{194}$$

$$\mathsf{g}_1(\mathsf{y}) = b\mathsf{y},\ \ b = \frac{a\sigma_{\mathsf{S}}^2}{a^2\sigma_{\mathsf{S}}^2 + \sigma_{\mathsf{N}}^2} \tag{195}$$

is an (n, d, ϵ, P) symbol-by-symbol code (with average cost constraint) such that

$$\mathbb{P}\left[W_0^n D(C(P)) > nd\right] = \epsilon \tag{196}$$

where W_0^n is chi-square distributed with n degrees of freedom.

Note that (196) is a particularization of (179). Using (196), we find that

$$\mathcal{W}_1(P) = 2D^2(C(P)) \log^2 e \tag{197}$$

On the other hand, using (122), we compute

$$\mathcal{W}(1, P) = D^2(C(P)) \left(2 - D^2(C(P))\right) \log^2 e \tag{198}$$

which means that for $\sigma_{\mathsf{S}}^2 < 1 + P$

$$\mathcal{W}_1(P) > \mathcal{W}(1, P) \tag{199}$$

The difference between (199) and (186) is due to the fact that the optimal symbol-by-symbol code in Theorem 23 obeys an average power constraint, rather than the more stringent maximal power constraint of Theorem 9, so it is not surprising that for $\epsilon > \frac{1}{2}$ the symbol-by-symbol code outperforms the best code obeying the maximal power constraint. More interestingly, in the practically relevant case $\epsilon < \frac{1}{2}$, (199) implies that the symbol-by-symbol code of Theorem 23 is suboptimal in terms of dispersion, even though it achieves the minimum average distortion. Nevertheless, in the range of blocklenghts displayed in Figure 10, the symbol-by-symbol code even outperforms the converse for codes operating under a maximal power constraint.

IX. Conclusion

The approach taken in this paper to analyze the non-asymptotic fundamental limits of lossy joint source-channel coding is two-fold. Our new achievability and converse bounds apply to abstract sources and channels and allow for memory, while the asymptotic analysis of the new

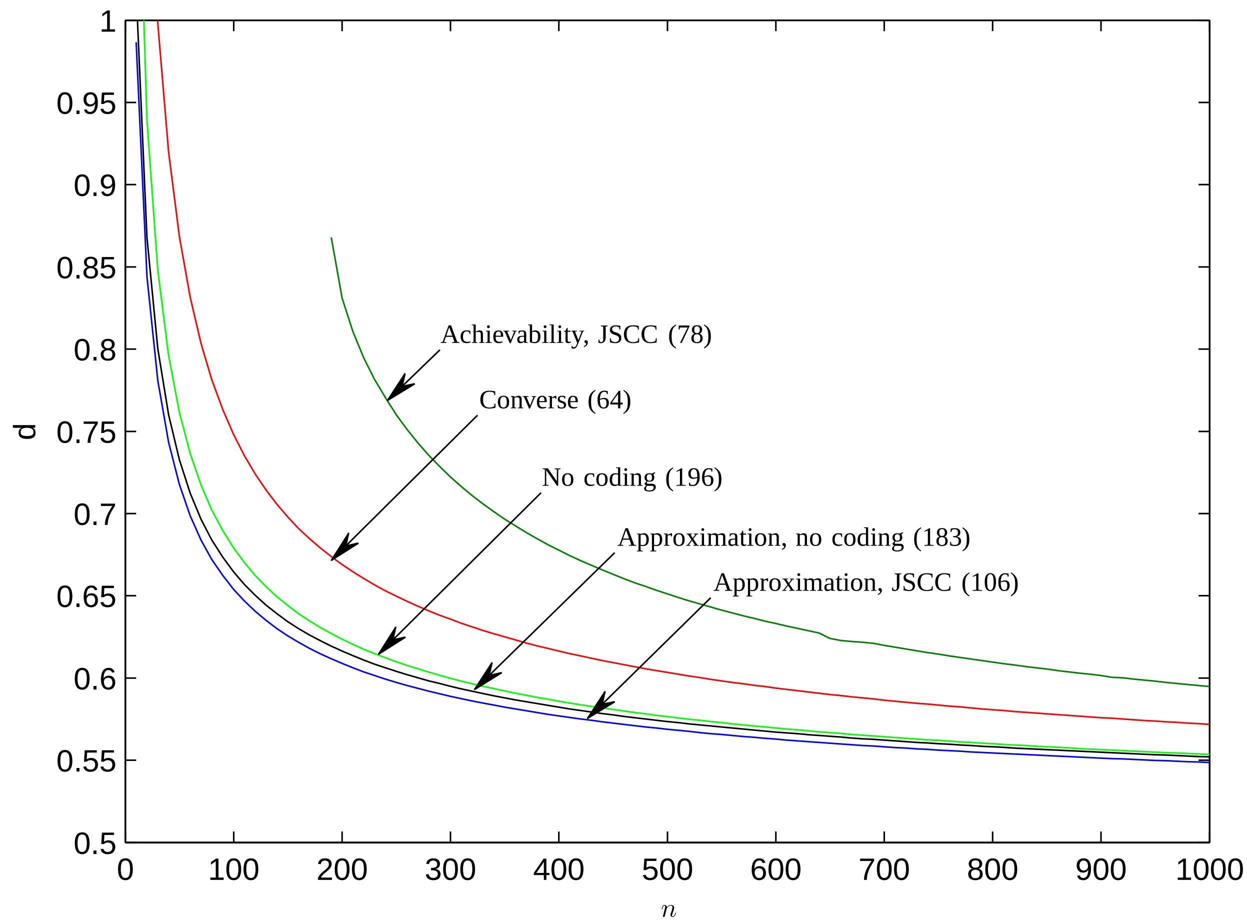


Fig. 10. Distortion-blocklength tradeoff for the transmission of a GMS over an AWGN channel with $\frac{P}{\sigma_{\mathrm{N}}^2} = 1$ and $R = 1$, $\epsilon = 10^{-2}$.

bounds leading to the dispersion of JSCC is focused on the most basic scenario of transmitting a stationary memoryless source over a stationary memoryless channel.

The major results and conclusions are the following.

- A general new converse bound (Theorem 3) leverages the concept of d-tilted information (Definition 5), a random variable which corresponds (in a sense that can be formalized [12], [22]) to the number of bits required to represent a given source outcome within distortion d and whose role in lossy compression is on a par with that of information (in (16)) in lossless compression.
- The converse result in Theorem 4 capitalizes on two simple observations, namely, that any (d, ϵ) lossy code can be converted to a list code with list error probability ϵ, and that a

binary hypothesis test between P_{SXY} and an auxiliary distribution on the same space can be constructed by choosing P_{SXY} when there is no list error.

- As evidenced by our numerical results, the converse result in Theorem 5, which applies to those channels satisfying a certain symmetry condition and which is a consequence of the hypothesis testing converse in Theorem 4, can outperform the d-tilted information converse in Theorem 3. Nevertheless, it is Theorem 3 that lends itself to analysis more easily and that leads to the JSCC dispersion for the general DMC.
- Our random-coding-based achievability bound (Theorem 7) provides insights into the degree of separation between the source and the channel coding required for optimal performance in the finite blocklength regime. More precisely, it reveals that the dispersion of JSCC can be achieved in the class of (M, d, ϵ) JSCC codes (Definition 8). As in separate source/channel coding, in (M, d, ϵ) coding the inner channel coding block is connected to the outer source coding block by a noiseless link of capacity $\log M$, but unlike SSCC, the channel (resp. source) code can be chosen based on the knowledge of the source (resp. channel). The conventional SSCC in which the source code is chosen without knowledge of the channel and the channel code is chosen without knowledge of the source, although known to achieve the asymptotic fundamental limit of joint source-channel coding under certain quite general conditions, is in general suboptimal in the finite blocklength regime.
- For the transmission of a stationary memoryless source over a stationary memoryless channel, the Gaussian approximation in Theorem 9 provides a simple estimate of the maximal nonasymptotically achievable joint source-channel coding rate. Appealingly, the dispersion of joint source-channel coding decomposes into two terms, the channel dispersion and the source dispersion. Thus, only two channel attributes, the capacity and dispersion, and two source attributes, the rate-distortion and rate-dispersion functions, are required to compute the Gaussian approximation to the maximal JSCC rate.
- In those curious cases when the source and the channel are probabilistically matched so that symbol-by-symbol coding attains the minimum possible average distortion, Theorem 21 ensures that it also attains the dispersion of joint source-channel coding, that is, symbol-by-symbol coding results in the minimum variance of distortions among all codes operating at that average distortion.
- Even in the absence of a probabilistic match between the source and the channel, symbol-by-

symbol transmission, though asymptotically suboptimal, might outperform separate source-channel coding and joint source-channel random coding in the finite blocklength regime.

APPENDIX A
THE BERRY-ESSEEN THEOREM

The following result is an important tool in the Gaussian approximation analysis.

Theorem 24 (Berry-Esseen CLT, e.g. [23, Ch. XVI.5 Theorem 2])**.** *Fix a positive integer* n*. Let* W_i*,* $i = 1, \ldots, n$ *be independent. Then, for any real* t

$$\left| \mathbb{P}\left[\sum_{i=1}^{n} W_i > n \left(D_n + t\sqrt{\frac{V_n}{n}} \right) \right] - Q(t) \right| \leq \frac{B_n}{\sqrt{n}}, \tag{200}$$

where

$$D_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[W_i\right] \tag{201}$$

$$V_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}\left[W_i\right] \tag{202}$$

$$T_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left| W_i - \mathbb{E}\left[W_i\right] \right|^3 \right] \tag{203}$$

$$B_n = \frac{c_0 T_n}{V_n^{3/2}} \tag{204}$$

and $0.4097 \leq c_0 \leq 0.5600$ *(*$0.4097 \leq c_0 < 0.4784$ *for identically distributed* W_i*).*

APPENDIX B
AUXILIARY RESULT ON THE MINIMIZATION OF THE INFORMATION SPECTRUM

Given a finite set $\mathcal{A}$, we say that $x^n \in \mathcal{A}^n$ has type P_{X} if the number of times each letter $a \in \mathcal{A}$ is encountered in x^n is $nP_{\mathsf{X}}(a)$. Let $\mathcal{P}$ be the set of all distributions on $\mathcal{A}$, which is simply the standard $|\mathcal{A}| - 1$ simplex in $\mathbb{R}^{|\mathcal{A}|}$. For an arbitrary subset $\mathcal{D} \subseteq \mathcal{P}$, denote by $\mathcal{D}_{[n]}$ the set of distributions in $\mathcal{D}$ that are also n-types, that is,

$$\mathcal{D}_{[n]} = \left\{ P_{\mathsf{X}} \in \mathcal{D} \colon \exists x^n \in \mathcal{A}^n \colon \mathrm{type}(x^n) = P_{\mathsf{X}} \right\} \tag{205}$$

Denote by $\Pi(P_{\mathsf{X}})$ the minimum Euclidean distance approximation of P_{X} in the set of n-types, that is,

$$\Pi(P_{\mathsf{X}}) = \arg \min_{\hat{P}_{\mathsf{X}} \in \mathcal{P}_{[n]}} \left| P_{\mathsf{X}} - \hat{P}_{\mathsf{X}} \right| \tag{206}$$

Let $\mathcal{P}^\star$ be the set of capacity-achieving distributions:

$$\mathcal{P}^\star = \{P_{\mathsf{X}} \in \mathcal{P} \colon I(\mathsf{X};\mathsf{Y}) = C\} \tag{207}$$

Denote the minimum (maximum) information variances achieved by the distributions in $\mathcal{P}^\star$ by

$$V_{\min} = \min_{P_{\mathsf{X}} \in \mathcal{P}^\star} \mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y})\right] \tag{208}$$

$$V_{\max} = \max_{P_{\mathsf{X}} \in \mathcal{P}^\star} \mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y})\right] \tag{209}$$

and let $\mathcal{P}^\star_{\min} \subseteq \mathcal{P}^\star$ be the set of capacity-achieving distributions that achieve the minimum (maximum) information variance:

$$\mathcal{P}^\star_{\min} = \{P_{\mathsf{X}} \in \mathcal{P}^\star \colon \mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y})\right] = V_{\min}\} \tag{210}$$

and analogously $\mathcal{P}^\star_{\max}$ for the distributions in $\mathcal{P}^\star$ with maximal variance. Lemma 1 below allows to show that in the memoryless case, the infimum inside the expectation in (37) is approximately attained by those sequences whose type is closest to the capacity-achieving distribution $P_{\mathsf{X}^\star}$ (if it is non-unique, $P_{\mathsf{X}^\star}$ is chosen appropriately based on the information variance it achieves). This technical result is the key to proving the converse part of Theorem 9.

Lemma 1. *There exist $\bar{\Delta} > 0$ such that for all sufficiently large n:*

1) *If $V_{\min} > 0$, then there exists $K > 0$ such that for $|\Delta| \le \bar{\Delta}$,*

$$\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i;Y_i) \le n\left(C - \Delta\right)\right] \ge \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i^\star;Y_i) \le n\left(C - \Delta\right)\right] - \frac{K}{\sqrt{n}} \tag{211}$$

where (211) *holds for any $x^{n\star}$ whose type is in $\mathcal{P}^\star_{\min}$ if $\Delta \ge 0$, and for any $x^{n\star}$ whose type is in $\mathcal{P}^\star_{\max}$ if $\Delta < 0$.*

2) *If $V_{\max} = 0$, then for all $0 < \alpha < \frac{3}{2}$ and $\Delta > \frac{\bar{\Delta}}{n^{\frac{1}{2}+\alpha}}$,*

$$\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i;Y_i) \le n\left(C + \Delta\right)\right] \ge 1 - \frac{1}{n^{\frac{1}{4} - \frac{3}{2}\alpha}} \tag{212}$$

The information densities in the left sides of (211) *and* (212) *are computed with* $\mathrm{type}(x^n) = P_{\mathsf{X}} \to P_{\mathsf{Y}|\mathsf{X}} \to P_{\mathsf{Y}}$*, and that in the right side of* (211) *is computed with* $\mathrm{type}(x^{n\star}) = P_{\mathsf{X}} \to P_{\mathsf{Y}|\mathsf{X}} \to P_{\mathsf{Y}}$*. The independent random variables Y_i in the left sides of* (211) *and* (212) *have distribution $P_{\mathsf{Y}|\mathsf{X}=x_i}$, while Y_i in the right side of* (211) *have distribution $P_{\mathsf{Y}|\mathsf{X}=x_i^\star}$.*

In order to prove Lemma 1, we first show three auxiliary lemmas. The first two deal with approximate optimization of functions.

If f and g approximate each other, and the minimum of f is approximately attained at x, then g is also approximately minimized at x, as the following lemma formalizes.

Lemma 2. *Fix $\eta > 0$, $\xi > 0$. Let $\mathcal{D}$ be an arbitrary set, and let $f\colon \mathcal{D} \mapsto \mathbb{R}$ and $g\colon \mathcal{D} \mapsto \mathbb{R}$ be such that*

$$\sup_{x \in \mathcal{D}} |f(x) - g(x)| \leq \eta \tag{213}$$

Further, assume that f and g attain their minima. Then,

$$g(x) \leq \min_{y \in \mathcal{D}} g(y) + \xi + 2\eta \tag{214}$$

as long as x satisfies

$$f(x) \leq \min_{y \in \mathcal{D}} f(y) + \xi \tag{215}$$

(see Fig. 11).

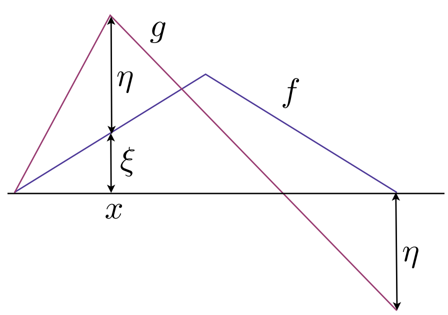


Fig. 11. An example where (214) holds with equality.

Proof of Lemma 2: Let $x^\star \in \mathcal{D}$ be such that $g(x^\star) = \min_{y \in \mathcal{D}} g(y)$. Using (213) and (215),

write

$$g(x) \leq \min_{y \in \mathcal{D}} f(y) + g(x) - f(x) + \xi \tag{216}$$

$$\leq \min_{y \in \mathcal{D}} f(y) + \eta + \xi \tag{217}$$

$$\leq f(x^\star) + \eta + \xi \tag{218}$$

$$= g(x^\star) - g(x^\star) + f(x^\star) + \eta + \xi \tag{219}$$

$$\leq g(x^\star) + 2\eta + \xi \tag{220}$$

■

The following lemma is reminiscent of [14, Lemma 64].

Lemma 3. *Let $\mathcal{D}$ be a compact metric space, and let $d\colon \mathcal{D}^2 \to \mathbb{R}^+$ be a metric. Fix $f\colon \mathcal{D} \mapsto \mathbb{R}$ and $g\colon \mathcal{D} \mapsto \mathbb{R}$. Let*

$$\mathcal{D}^\star = \left\{ x \in \mathcal{D}\colon f(x) = \max_{y \in \mathcal{D}} f(y) \right\} \tag{221}$$

Suppose that for some constants $\ell > 0, L > 0$, we have, for all $(x, x^\star) \in \mathcal{D} \times \mathcal{D}^\star$,

$$f(x^\star) - f(x) \geq \ell d^2(x, x^\star) \tag{222}$$

$$|g(x^\star) - g(x)| \leq L d(x, x^\star) \tag{223}$$

Then, for any positive scalars φ, ψ,

$$\max_{x \in \mathcal{D}} \left[\varphi f(x) \pm \psi g(x)\right] \leq \varphi f(x^\star) \pm \psi g(x^\star) + \frac{L^2 \psi^2}{4 \ell \varphi} \tag{224}$$

Moreover, if, instead of (222), *f satisfies*

$$f(x^\star) - f(x) \geq \ell d(x, x^\star) \tag{225}$$

then, for any positive scalars ψ, φ such that

$$L\psi \leq \ell\varphi \tag{226}$$

we have

$$\max_{x \in \mathcal{D}} \left[\varphi f(x) \pm \psi g(x)\right] \leq \varphi f(x^\star) \pm \psi g(x^\star) \tag{227}$$

Proof of Lemma 3: Let x_0 achieve the maximum on the left side of (224). Using (222) and (223), we have, for all $x^\star \in \mathcal{D}^\star$,

$$0 \leq \varphi\left(f(x_0) - f(x^\star)\right) \pm \psi\left(g(x_0) - g(x^\star)\right) \tag{228}$$

$$\leq -\ell\varphi d^2(x_0, x^\star) + L\psi d(x_0, x^\star) \tag{229}$$

$$\leq \frac{L^2\psi^2}{4\ell\varphi} \tag{230}$$

where (230) follows because the maximum of (229) is achieved at $d(x_0, x^\star) = \frac{L\psi}{2\ell\varphi}$.

To show (227), observe using (225) and (223) that

$$0 \leq \varphi\left(f(x_0) - f(x^\star)\right) \pm \psi\left(g(x_0) - g(x^\star)\right) \tag{231}$$

$$\leq \left(-\ell\varphi + L\psi\right) d(x_0, x^\star) \tag{232}$$

$$\leq 0 \tag{233}$$

where where (233) follows from (226).

■

The following lemma deals with asymptotic behavior of the Q-function.

Lemma 4. *Fix $a \geq 0$, $b \geq 0$. Then, there exists $q \geq 0$ (explicitly computed in the proof) such that for all $z \geq -\frac{\sqrt{n}}{b}$ and all n large enough,*

$$Q\left(z - \frac{a}{\sqrt{n}}\right) - Q\left(z + \frac{b}{\sqrt{n}}z^2\right) \leq \frac{q}{\sqrt{n}} \tag{234}$$

Proof of Lemma 4:

$Q(x)$ is convex for $x \geq 0$, and $Q'(x) = -\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, so for $x \geq 0$, $\xi \geq 0$

$$Q(x + \xi) \geq Q(x) - \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\xi \tag{235}$$

while for arbitrary x and $\xi \geq 0$,

$$Q(x + \xi) \geq Q(x) - \frac{1}{\sqrt{2\pi}}\xi \tag{236}$$

If $z \geq \frac{a}{\sqrt{n}}$, we use (235) to obtain

$$Q\left(z-\frac{a}{\sqrt{n}}\right)-Q\left(z+\frac{b}{\sqrt{n}}z^2\right) \tag{237}$$

$$\leq \frac{1}{\sqrt{2\pi}}e^{-\frac{\left(z-\frac{a}{\sqrt{n}}\right)^2}{2}}\left(\frac{b}{\sqrt{n}}z^2+\frac{a}{\sqrt{n}}\right) \tag{238}$$

$$\leq \frac{bz^2}{\sqrt{2\pi n}}e^{-\frac{\left(z-\frac{a}{\sqrt{n}}\right)^2}{2}}+\frac{a}{\sqrt{2\pi n}} \tag{239}$$

$$\leq \frac{3b\,e^{-\frac{1}{2}}+a}{\sqrt{2\pi n}} \tag{240}$$

where (240) holds for n large enough because the maximum of (239) is attained at $z = \sqrt{2+\frac{a}{2n}}+\frac{a}{2\sqrt{n}}$.

If $0 \leq z \leq \frac{a}{\sqrt{n}}$, we use (236) to obtain

$$Q\left(z-\frac{a}{\sqrt{n}}\right)-Q\left(z+\frac{b}{\sqrt{n}}z^2\right) \leq \frac{1}{\sqrt{2\pi}}\left(\frac{b}{\sqrt{n}}z^2+\frac{a}{\sqrt{n}}\right) \tag{241}$$

$$\leq \frac{a}{\sqrt{2\pi n}}\left(1+\frac{ab}{n}\right) \tag{242}$$

If $-\frac{\sqrt{n}}{b} \leq z \leq 0$, we use $Q(x) = 1-Q(-x)$ to obtain

$$Q\left(z-\frac{a}{\sqrt{n}}\right)-Q\left(z+\frac{b}{\sqrt{n}}z^2\right)$$

$$= Q\left(|z|-\frac{b}{\sqrt{n}}z^2\right)-Q\left(|z|+\frac{a}{\sqrt{n}}\right) \tag{243}$$

$$\leq \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2\left(1-\frac{b}{\sqrt{n}}|z|\right)^2}{2}}\left(\frac{b}{\sqrt{n}}z^2+\frac{a}{\sqrt{n}}\right) \tag{244}$$

$$\leq \frac{bz^2}{\sqrt{2\pi n}}e^{-\frac{z^2\left(1-\frac{b}{\sqrt{n}}|z|\right)^2}{2}}+\frac{a}{\sqrt{2\pi n}} \tag{245}$$

$$\leq \frac{bz^2}{\sqrt{2\pi n}}e^{-\frac{z^2}{4}}+\frac{a}{\sqrt{2\pi n}} \tag{246}$$

$$\leq \frac{4be^{-1}+a}{\sqrt{2\pi n}} \tag{247}$$

where to justify (246), which holds for n large enough, observe that for such n the function in (246) is monotonically decreasing for $|z| \geq \sqrt{3}$, so the maximum of (246) is attained at some $|z| < \sqrt{3}$. But for such $|z|$, we may lower bound $\left(1-\frac{b}{\sqrt{n}}|z|\right)^2 \geq \frac{1}{2}$ in (245) for sufficiently large n, and (246) follows. ■

We are now equipped to prove Lemma 1.

Proof of Lemma 1:

Define the following functions $\mathcal{P} \mapsto \mathbb{R}_+$:

$$I(P_{\mathsf{X}}) = I(\mathsf{X}; \mathsf{Y}) = \mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X}; \mathsf{Y})\right] \tag{248}$$

$$V(P_{\mathsf{X}}) = \mathbb{E}\left[\mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X}; \mathsf{Y}) \mid \mathsf{X}\right]\right] \tag{249}$$

$$T(P_{\mathsf{X}}) = \mathbb{E}\left[\left|\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X}; \mathsf{Y}) - \mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X}; \mathsf{Y})|\mathsf{X}\right]\right|^3 \mid \mathsf{X}\right] \tag{250}$$

If $P_{\mathsf{X}} = \mathrm{type}(x^n)$, then for each $a \in \mathcal{A}$, there are $nP_{\mathsf{X}}(a)$ occurrences of $P_{\mathsf{Y}|\mathsf{X}=a}$ among the $\{P_{\mathsf{Y}|\mathsf{X}=x_i}, i = 1, 2, \ldots, n\}$, and in the sequel will invoke Theorem 24 with $W_i = \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i)$ where x^n is a given sequence, and (201)–(203) become

$$D_n = \frac{1}{n} \sum_{a=1}^{|\mathcal{A}|} nP_{\mathsf{X}}(a) \mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}}(a; \mathsf{Y}) \mid \mathsf{X} = a\right] \tag{251}$$

$$= I(P_{\mathsf{X}}) \tag{252}$$

$$V_n = \frac{1}{n} \left[\sum_{a=1}^{|\mathcal{A}|} nP_{\mathsf{X}}(a) \mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}}(a; \mathsf{Y}) \mid \mathsf{X} = a\right]\right] \tag{253}$$

$$= V(P_{\mathsf{X}}) \tag{254}$$

$$T_n = \frac{1}{n} \left[\sum_{a=1}^{|\mathcal{A}|} nP_{\mathsf{X}}(a) \left|\imath_{\mathsf{X};\mathsf{Y}}(a; \mathsf{Y}) - \mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}}(a; \mathsf{Y})|\mathsf{X} = a\right]\right|^3\right] \tag{255}$$

$$= T(P_{\mathsf{X}}) \tag{256}$$

Define the (Euclidean) δ-neighborhood of the set of capacity-achieving distributions $\mathcal{P}^\star$,

$$\mathcal{P}^\star_\delta = \left\{P_{\mathsf{X}} \in \mathcal{P} \colon \min_{P_{\mathsf{X}^\star} \in \mathcal{P}^\star} |P_{\mathsf{X}} - P_{\mathsf{X}^\star}| \leq \delta\right\} \tag{257}$$

We split the domain of the minimization in the right side of (211) into two sets, $\mathrm{type}(x^n) \in \mathcal{P}^\star_{\delta,[n]}$ and $\mathrm{type}(x^n) \in \mathcal{P}_{[n]} \backslash \mathcal{P}^\star_\delta$ (recall notation (205)), for an appropriately chosen $\delta > 0$.

We now show that (211) holds for all $\Delta \leq \frac{\Delta_I}{2}$ if the minimization is restricted to types in $\mathcal{P}_{[n]} \backslash \mathcal{P}^\star_\delta$, where $\delta > 0$ is arbitrary, and

$$\Delta_I = C - \max_{P_{\mathsf{X}} \in \mathcal{P}_{[n]} \backslash \mathcal{P}^\star_\delta} I(P_{\mathsf{X}}) > 0 \tag{258}$$

By Chebyshev's inequality, for all x^n whose type belongs to $\mathcal{P}_{[n]} \backslash \mathcal{P}_\delta^\star$,

$$\mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) > n(C - \Delta)\right] \tag{259}$$

$$= \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) - nI(P_\mathsf{X}) > n(C - I(P_\mathsf{X})) - n\Delta\right] \tag{260}$$

$$\leq \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) - nI(P_\mathsf{X}) > \frac{n\Delta_I}{2}\right] \tag{261}$$

$$\leq \mathbb{P}\left[\left(\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) - nI(P_\mathsf{X})\right)^2 > \frac{n^2\Delta_I^2}{4}\right] \tag{262}$$

$$\leq \frac{4nV(P_\mathsf{X})}{n^2\Delta_I^2} \tag{263}$$

$$\leq \frac{4\overline{V}}{n\Delta_I^2} \tag{264}$$

where in (261) we used

$$\Delta \leq \frac{1}{2}\Delta_I < \Delta_I \leq C - I(P_\mathsf{X}) \tag{265}$$

and

$$\overline{V} = \max_{P_\mathsf{X} \in \mathcal{P}} V(P_\mathsf{X}) \tag{266}$$

Note that $\overline{V} < \infty$ by Property 1 below. Therefore,

$$\min_{\mathrm{type}(x^n) \in \mathcal{P}_{[n]} \backslash \mathcal{P}_\delta^\star} \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \leq n(C - \Delta)\right] \tag{267}$$

$$> 1 - \frac{4\overline{V}}{n\Delta_I^2} \tag{268}$$

$$\geq \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i^\star; Y_i) \leq n(C - \Delta)\right] - \frac{4\overline{V}}{n\Delta_I^2} \tag{269}$$

We conclude that (211) holds if the minimization is restricted to types in $\mathcal{P}_{[n]} \backslash \mathcal{P}_\delta^\star$.

Without loss of generality, we assume that all outputs in $\mathcal{B}$ are accessible (which implies that $P_{\mathsf{Y}^\star}(\mathsf{y}) > 0$ for all $\mathsf{y} \in \mathcal{B}$) and choose $\delta > 0$ so that for all $P_\mathsf{X} \in \mathcal{P}_\delta^\star$ and $\mathsf{y} \in \mathcal{B}$,

$$P_\mathsf{Y}(\mathsf{y}) > 0 \tag{270}$$

We recall the following properties of the functions $I(\cdot)$, $V(\cdot)$ and $T(\cdot)$ from [14, Appendices E and I].

Property 1. *The functions $I(P_{\mathsf{X}})$, $V(P_{\mathsf{X}})$ and $T(P_{\mathsf{X}})$ are continuous on the compact set $\mathcal{P}$, and therefore bounded and achieve their extrema.*

Property 2. *There exists an $\ell_1 > 0$ such that for all $(P_{\mathsf{X}^\star}, P_{\mathsf{X}}) \in \mathcal{P}^\star \times \mathcal{P}_\delta^\star$,*

$$C - I(P_{\mathsf{X}}) \geq \ell_1 \left|P_{\mathsf{X}} - P_{\mathsf{X}^\star}\right|^2 \tag{271}$$

Property 3. *In $\mathcal{P}_\delta^\star$, the functions $I(P_{\mathsf{X}})$, $V(P_{\mathsf{X}})$ and $T(P_{\mathsf{X}})$ are infinitely differentiable.*

Property 4. *In $\mathcal{P}^\star$, $V(P_{\mathsf{X}}) = \mathrm{Var}\left[\imath_{\mathsf{X};\mathsf{Y}}(\mathsf{X};\mathsf{Y})\right]$.*

Due to Property 3, there exist nonnegative constants L_1 and L_2 such that for all $(P_{\mathsf{X}}, P_{\mathsf{X}^\star}) \in \mathcal{P}_\delta^\star \times \mathcal{P}^\star$,

$$C - I(P_{\mathsf{X}}) \leq L_1 \left|P_{\mathsf{X}} - P_{\mathsf{X}^\star}\right| \tag{272}$$

$$V(P_{\mathsf{X}}) \leq L_2 \left|P_{\mathsf{X}} - P_{\mathsf{X}^\star}\right| \tag{273}$$

To treat the case $x^n \in \mathcal{P}^\star_{\delta,[n]}$, we will need to choose $\delta > 0$ carefully and to consider the cases $V_{\min} > 0$ and $V_{\max} = 0$ separately.

A. $V_{\min} > 0$.

We decrease δ until

$$V_{\min} \leq 2 \min_{P_{\mathsf{X}} \in \mathcal{P}_\delta^\star} V\left(P_{\mathsf{X}}\right) \tag{274}$$

$$\delta \leq \min\left\{ \frac{V_{\min}^{\frac{3}{2}}}{4L_2\sqrt{2\overline{V}}},\ \frac{L_2}{2\sqrt{V_{\min}}},\ \frac{\bar{\Delta}}{4H} \right\} \tag{275}$$

are satisfied, in addition to (270), where $\overline{V}$, L_1, L_2 are defined in (266), (272) and (295), respectively.

We now show that (211) holds if the minimization is restrained to types in $\mathcal{P}^\star_{\delta,[n]}$, for all $\Delta \geq -\underline{\Delta}$, for an appropriately chosen $\underline{\Delta} > 0$. Using (274) and boundedness of $T(P_{\mathsf{X}})$, write

$$B = \max_{P_{\mathsf{X}} \in \mathcal{P}_\delta^\star} \frac{c_0 T(P_{\mathsf{X}})}{V^{\frac{3}{2}}(P_{\mathsf{X}})} \leq \frac{2^{\frac{3}{2}} c_0 \overline{T}}{V_{\min}^{\frac{3}{2}}} < \infty \tag{276}$$

where

$$\overline{T} = \max_{P_{\mathsf{X}} \in \mathcal{P}^\star_\delta} T(P_{\mathsf{X}}) < \infty \tag{277}$$

Therefore, for any x^n with $\mathrm{type}(x^n) \in \mathcal{P}^\star_{\delta,[n]}$, the Berry-Esseen bound yields:

$$\left| \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \leq n(C - \Delta) \right] - Q\left(t(P_{\mathsf{X}})\right) \right| \leq \frac{B}{\sqrt{n}} \tag{278}$$

where

$$\nu(P_{\mathsf{X}}) = \frac{nI(P_{\mathsf{X}}) - nC + n\Delta}{\sqrt{nV(P_{\mathsf{X}})}} \tag{279}$$

We now apply Lemma 2 with $\mathcal{D} = \mathcal{P}^\star_{\delta,[n]}$ and

$$f(P_{\mathsf{X}}) = Q\left(\nu(P_{\mathsf{X}})\right) \tag{280}$$

$$g(P_{\mathsf{X}}) = \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \leq n(C - \Delta) \right] \tag{281}$$

Condition (213) of Lemma 2 holds with $\eta = \frac{B}{\sqrt{n}}$ due to (278). As will be shown in the sequel, the following version of condition (215) holds:

$$Q(\nu(\Pi(P_{\mathsf{X}^\star}))) \leq \min_{P_{\mathsf{X}} \in \mathcal{P}^\star_{\delta,[n]}} Q(\nu(P_{\mathsf{X}})) + \frac{q}{\sqrt{n}} \tag{282}$$

where $\Pi(P_{\mathsf{X}})$, the minimum Euclidean distance approximation of P_{X} in the set of n-types, is formally defined in (206), and $q > 0$ will be chosen later. We proceed to apply Lemma 2, and (214) leads to

$$\begin{aligned} &\min_{\mathrm{type}(x^n) \in \mathcal{P}^\star_{\delta,[n]}} \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \leq n(C - \Delta) \right] \\ &\qquad \geq \mathbb{P}\left[\sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i^\star; Y_i) \leq n(C - \Delta) \right] - \frac{q + 2B}{\sqrt{n}} \end{aligned} \tag{283}$$

We conclude that (211) holds if minimization is restrained to types in $\mathcal{P}^\star_{\delta,[n]}$.

We proceed to show (282). As will be proven later, for appropriately chosen $\underline{L} > 0$ and $L > 0$

it holds that

$$\frac{\sqrt{n}\Delta}{\sqrt{V(P_{\mathsf{X}^\star})}} - \frac{L}{2}\sqrt{\frac{|\mathcal{A}|}{n}} \le \nu(\Pi\left(P_{\mathsf{X}^\star}\right)) \tag{284}$$

$$\le \max_{P_{\mathsf{X}} \in \mathcal{P}^\star_{\delta,[n]}} \nu(P_{\mathsf{X}}) \tag{285}$$

$$\le \max_{P_{\mathsf{X}} \in \mathcal{P}^\star_{\delta}} \nu(P_{\mathsf{X}}) \tag{286}$$

$$\le \frac{\sqrt{n}\Delta}{\sqrt{V(P_{\mathsf{X}^\star})}} + \frac{\sqrt{n}L^2\Delta^2}{4\ell} \tag{287}$$

where $P_{\mathsf{X}^\star} \in \mathcal{P}^\star_{\min}$ if $\Delta \ge 0$ and $P_{\mathsf{X}^\star} \in \mathcal{P}^\star_{\max}$ if $\Delta < 0$.

Denote

$$a = \frac{L\sqrt{|\mathcal{A}|}}{2} \tag{288}$$

$$b = \frac{V(P_{\mathsf{X}^\star})L^2}{4K} \tag{289}$$

$$z = \frac{\sqrt{n}\Delta}{\sqrt{V(P_{\mathsf{X}^\star})}} \tag{290}$$

If

$$\Delta \ge -\frac{4\ell}{L^2\sqrt{V_{\max}}} = -\underline{\Delta} \tag{291}$$

then $z \ge -\frac{\sqrt{n}}{b}$, and Lemma 4 applies to z. So, using (284), (287), the fact that $Q(\cdot)$ is monotonically decreasing and Lemma 4, we conclude that there exists a $q > 0$ such that

$$Q\left(\nu(\Pi(P_{\mathsf{X}^\star}))\right) - \min_{P_{\mathsf{X}} \in \mathcal{P}^\star_{\delta,[n]}} Q(\nu(P_{\mathsf{X}}))$$

$$= Q\left(\nu(\Pi(P_{\mathsf{X}^\star}))\right) - Q\left(\max_{P_{\mathsf{X}} \in \mathcal{P}^\star_{\delta,[n]}} \nu(P_{\mathsf{X}})\right) \tag{292}$$

$$\le Q\left(z - \frac{a}{\sqrt{n}}\right) - Q\left(z + \frac{b}{\sqrt{n}}z^2\right) \tag{293}$$

$$\le \frac{q}{\sqrt{n}} \tag{294}$$

which is equivalent to (282).

It remains to prove (284) and (287). Using (274), observe that in $\mathcal{P}^\star_\delta$, the gradient with respect to P_{X} satisfies

$$\left|\nabla \frac{1}{\sqrt{V(P_{\mathsf{X}})}}\right| \le L = \frac{L_2\sqrt{2}}{V_{\min}^{\frac{3}{2}}} \tag{295}$$

so recalling (272) and denoting $\zeta = |P_{\mathsf{X}} - P_{\mathsf{X}^\star}|$, we have

$$\frac{C - P_{\mathsf{X}} - \Delta}{\sqrt{V(P_{\mathsf{X}})}} \le -\frac{\Delta}{\sqrt{V(P_{\mathsf{X}^\star})}} + \frac{L_1\zeta}{\sqrt{V(P_{\mathsf{X}^\star})}} + L\zeta\left(\underline{\Delta} + L_1\zeta\right) \tag{296}$$

$$\le -\frac{\Delta}{\sqrt{V(P_{\mathsf{X}^\star})}} + \underline{L}\zeta \tag{297}$$

where

$$\underline{L} = \frac{L_1}{\sqrt{V_{\min}}} + L\underline{\Delta} + LL_1\delta \tag{298}$$

So, (284) follows by observing

$$|P_{\mathsf{X}} - \Pi\left(P_{\mathsf{X}}\right)| \le \frac{\sqrt{|\mathcal{A}|}}{2n} \tag{299}$$

To show (287), we apply Lemma 3 with

$$\mathcal{D} = \mathcal{P}^\star_\delta \tag{300}$$

$$\mathcal{D}^\star = \mathcal{P}^\star \tag{301}$$

$$\varphi = \sqrt{n} \tag{302}$$

$$\psi = \sqrt{n}|\Delta| \tag{303}$$

$$f\left(P_{\mathsf{X}}\right) = \frac{I(P_{\mathsf{X}}) - C}{\sqrt{V(P_{\mathsf{X}})}} \tag{304}$$

$$g\left(P_{\mathsf{X}}\right) = \frac{1}{\sqrt{V(P_{\mathsf{X}})}} \tag{305}$$

Let us verify that conditions of Lemma 3 hold. Function g satisfies condition (223) with L defined in (295). Let us now show that function f satisfies condition (222) with $\ell = \frac{\ell_1}{2\sqrt{\overline{V}}}$, where $\overline{V}$ and ℓ_1 were defined in (266) and (271), respectively. Write

$$\max_{P_{\mathsf{X}} \in \mathcal{P}^\star_\delta} f(P_{\mathsf{X}}) - f(P_{\mathsf{X}}) = \frac{C - I(P_{\mathsf{X}})}{\sqrt{V(P_{\mathsf{X}})}} \tag{306}$$

$$= \left(C - I(P_{\mathsf{X}})\right) g(P_{\mathsf{X}}) \tag{307}$$

$$\ge \left(C - I(P_{\mathsf{X}})\right) \left(\max_{P_{\mathsf{X}} \in \mathcal{P}^\star_\delta} g(P_{\mathsf{X}}) - 2L\delta\right) \tag{308}$$

$$\ge \left(C - I(P_{\mathsf{X}})\right) \left(\frac{1}{\sqrt{\overline{V}}} - 2L\delta\right) \tag{309}$$

$$\ge \left(C - I(P_{\mathsf{X}})\right) \frac{1}{2\sqrt{\overline{V}}} \tag{310}$$

$$\ge \frac{\ell_1}{2\sqrt{\overline{V}}} \left|P_{\mathsf{X}} - P_{\mathsf{X}^\star}\right|^2 \tag{311}$$

where

- (308) follows from (223) and (257);
- (309) uses notation (266);
- (310) follows from (275);
- (311) applies (271).

So, Lemma 3 applies to $\nu(P_{\mathsf{X}}) = \varphi f(P_{\mathsf{X}}) + \operatorname{sign}(\Delta)\psi g(P_{\mathsf{X}})$, resulting in (287), thereby completing the proof of (283).

Combining (269) and (283), we conclude that (211) holds for all Δ in the interval

$$-\frac{4\ell}{L^2\sqrt{V_{\max}}} \le \Delta \le \frac{\Delta_I}{2} \tag{312}$$

B. $V_{\max} = 0$.

We choose δ so that (270) is satisfied. The case $\operatorname{type}(x^n) \notin \mathcal{P}^{\star}_{\delta,[n]}$ was covered in (269), so we only need to consider minimization of the left side of (212) over $\mathcal{P}^{\star}_{\delta,[n]}$. Fix $\alpha < \frac{3}{2}$. If

$$\Delta \le -\frac{3\ell_1 L^{\frac{2}{3}}}{(1+3\ell_1)^{\frac{4}{3}}} \frac{1}{n^{\frac{1}{2}+\alpha}} \tag{313}$$

we have

$$\mathbb{P}\left[\sum_{i=1}^{n} \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) > n(C-\Delta)\right] \tag{314}$$

$$\le \mathbb{P}\left[\sum_{i=1}^{n} \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) - nI(P_{\mathsf{X}}) > n(C - I(P_{\mathsf{X}})) + n|\Delta|\right] \tag{315}$$

$$\le \frac{V(P_{\mathsf{X}})}{n\left(C - I(P_{\mathsf{X}}) + |\Delta|\right)^2} \tag{316}$$

$$\le \frac{L|P_{\mathsf{X}} - P_{\mathsf{X}^\star}|}{n\left(\ell_1|P_{\mathsf{X}} - P_{\mathsf{X}^\star}|^2 + |\Delta|\right)^2} \tag{317}$$

$$\le \frac{L\left(3\ell_1\right)^{\frac{3}{2}}}{\left(1+3\ell_1\right)^2} \frac{1}{n|\Delta|^{\frac{3}{2}}} \tag{318}$$

$$\le \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} \tag{319}$$

where

- (316) is by Chebyshev's inequality;
- (317) uses (271), (295) and $V_{\max} = 0$;
- (318) holds because the maximum of its left side is attained at $|P_{\mathsf{X}} - P_{\mathsf{X}^\star}|^2 = \frac{|\Delta|}{3\ell_1}$.

■

APPENDIX C
PROOF OF THE CONVERSE PART OF THEOREM 9

Note that for the converse, restriction (iv) can be replaced by the following weaker one.

(iv′) The random variable $\jmath_{\mathsf{S}}(\mathsf{S}, d)$ has finite absolute third moment.

To verify that (iv) implies (iv′), observe that by the concavity of the logarithm,

$$0 \le \jmath_{\mathsf{S}}(\mathsf{s}, d) + \lambda^\star d \le \lambda^\star \mathbb{E}\left[\mathsf{d}(\mathsf{s}, \mathsf{Z}^\star)\right] \tag{320}$$

so

$$\mathbb{E}\left[\left|\jmath_{\mathsf{S}}(\mathsf{S}, d) + \lambda^\star d\right|^3\right] \le \lambda^{\star 3}\mathbb{E}\left[\mathsf{d}^3(\mathsf{S}, \mathsf{Z}^\star)\right] \tag{321}$$

We now proceed to prove the converse by showing first that we can eliminate all rates exceeding

$$\frac{k}{n} \ge \frac{C}{R(d) - 3\tau} \tag{322}$$

for an arbitrary $0 < \tau < \frac{R(d)}{3}$. More precisely, we show that the excess-distortion probability of any code having such rate converges to 1 as $n \to \infty$, and therefore for any $\epsilon < 1$, there is an n_0 such that for all $n \ge n_0$, no (k, n, d, ϵ) code can exist for k, n satisfying (322).

We weaken (20) by fixing $\gamma = k\tau$ and choosing a particular output distribution, namely, $P_{\bar{Y}^n} = P_{Y^{n\star}} = P_{\mathsf{Y}^\star} \times \ldots \times P_{\mathsf{Y}^\star}$. Due to (ii), $P^\star_{Z^k} = P^\star_{\mathsf{Z}} \times \ldots \times P^\star_{\mathsf{Z}}$, the $\mathsf{d}-$tilted information single-letterizes, that is, for a.e. s^k,

$$\jmath_{S^k}(s^k, d) = \sum_{i=1}^k \jmath_{\mathsf{S}}(s_i, d) \tag{323}$$

so Theorem 1 implies that the parameters of every (k, n, d, ϵ') code must satisfy

$$\epsilon' \ge \mathbb{E}\left[\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^k \jmath_{\mathsf{S}}(S_i, d) - \sum_{j=1}^n \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \ge k\tau \mid S^k\right]\right] - \exp\left(-k\tau\right) \tag{324}$$

$$\ge \min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{j=1}^n \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \le nC + k\tau\right] \cdot \mathbb{P}\left[\sum_{i=1}^k \jmath_{\mathsf{S}}(S_i, d) \ge nC + 2k\tau\right] - \exp\left(-k\tau\right) \tag{325}$$

$$\ge \min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{j=1}^n \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \le nC + n\tau'\right] \cdot \mathbb{P}\left[\sum_{i=1}^k \jmath_{\mathsf{S}}(S_i, d) \ge kR(d) - k\tau\right] - \exp\left(-k\tau\right) \tag{326}$$

where in (326), we used (322) and $\tau' = \frac{C\tau}{R(d)-3\tau} > 0$. Recalling (12) and

$$\mathbb{E}\left[\imath_{\mathsf{X};\mathsf{Y}^\star}(\mathsf{x};\mathsf{Y})|\mathsf{X}=\mathsf{x}\right] \le C \tag{327}$$

with equality for $P_{\mathsf{X}^\star}$-a.e. x, we conclude using the law of large numbers that (326) tends to 1 as $k, n \to \infty$.

We proceed to show that for all large enough k, n, if there is a sequence of (k, n, d, ϵ') codes such that

$$\begin{aligned} -3k\tau &\le nC - kR(d) && (328)\\ &\le \sqrt{nV + k\mathcal{V}(d)}Q^{-1}\left(\epsilon\right) + \theta\left(n\right) && (329) \end{aligned}$$

then $\epsilon' \ge \epsilon$.

Note that in general the bound in Theorem 1 does not lead to the correct channel dispersion term. We first consider the general case, in which we apply Theorem 3, and then we show the symmetric case, in which we apply Theorem 2.

Recall that $x^n \in \mathcal{A}^n$ has type P_{X} if the number of times each letter $a \in \mathcal{A}$ is encountered in x^n is $nP_{\mathsf{X}}(a)$. In Theorem 3, let t index the types P_{X} of sequences in $\mathcal{X} = \mathcal{A}^n$. Note that the total number of types satisfies [20] $T \le (n+1)^{|\mathcal{A}|-1}$. We will weaken the outer supremum in (37) by fixing a particular collection of output distributions, namely, $P_{Y^n_{\mathrm{type}(x^n)}} = P_{\mathsf{Y}} \times \ldots \times P_{\mathsf{Y}}$, where $P_{\mathsf{X}} \to P_{\mathsf{Y}|\mathsf{X}} \to P_{\mathsf{Y}}$, i.e. P_{Y} is the output distribution induced by $P_{\mathsf{X}} = \mathrm{type}(x^n)$. In this way, Theorem 3 implies that every (k, n, d, ϵ') code must satisfy

$$\begin{aligned} \epsilon' \ge \mathbb{E}&\left[\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^k \jmath_{\mathsf{S}}(S_i, d) - \sum_{i=1}^n \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \ge \gamma \mid S^k\right]\right] \\ &- (n+1)^{|\mathcal{A}|-1} \exp\left(-\gamma\right) && (330) \end{aligned}$$

Choose

$$\gamma = \left(|\mathcal{A}| - \frac{1}{2}\right)\log(n+1) \tag{331}$$

At this point we consider two cases separately, $V > 0$ and $V = 0$.

A. $V > 0$.

In order to apply Lemma 1 in Appendix B, we isolate the typical set of source sequences:

$$\mathcal{T}_{k,n} = \left\{ s^k \in \mathcal{S}^k \colon \left| \sum_{i=1}^k \jmath_{\mathsf{S}}(s_i, d) - nC - \gamma \right| \le n\bar{\Delta} \right\} \tag{332}$$

Observe that

$$\mathbb{P}\left[S^k \notin \mathcal{T}_{k,n}\right] = \mathbb{P}\left[\left|\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - nC\right| > n\bar{\Delta} - \gamma\right] \tag{333}$$

$$\leq \mathbb{P}\left[\left|\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - kR(d)\right| + |nC - kR(d)| + \gamma > n\bar{\Delta}\right] \tag{334}$$

$$\leq \mathbb{P}\left[\left|\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - kR(d)\right| > k\frac{\bar{\Delta}R(d)}{2C}\right] \tag{335}$$

$$\leq \frac{4C^2}{R^2(d)\bar{\Delta}^2}\frac{\mathcal{V}(d)}{k} \tag{336}$$

where

- (335) follows by lower bounding

$$n\bar{\Delta} - \gamma - |nC - kR(d)| \geq n\bar{\Delta} - \gamma - 3k\tau \tag{337}$$

$$\geq \frac{k\bar{\Delta}}{C}\left(R(d) - 3\tau\right) - \gamma - 3k\tau \tag{338}$$

$$\geq k\frac{\bar{\Delta}R(d)}{2C} \tag{339}$$

  where

  - (337) holds for large enough n due to (328) and (329);
  - (338) lower bounds n using (328);
  - (339) holds for a small enough $\tau > 0$.

- (336) is by Chebyshev's inequality.

Now, we let

$$\epsilon_{k,n} = \epsilon + \frac{B}{\sqrt{n+k}} + \frac{1}{\sqrt{n+1}} + \frac{4C^2}{R^2(d)\bar{\Delta}^2}\frac{\mathcal{V}(d)}{k} \tag{340}$$

where $B > 0$ will be chosen in the sequel, and k, n are chosen so that both (328) and the following version of (329) hold:

$$nC - kR(d) \leq \sqrt{nV + k\mathcal{V}(d) - \frac{L\sqrt{|\mathcal{A}|}}{2(n+k)}}Q^{-1}\left(\epsilon_{k,n}\right) - \gamma \tag{341}$$

where $L < \infty$ is the maximum absolute value of the gradient of the function $V(\cdot)$ (defined in (249)) over the set $P_\delta^\star$ (in (257)). Denote for brevity

$$r(x^n, y^n, s^n) = \sum_{i=1}^{n} \imath_{\mathsf{X};\mathsf{Y}}(x_i; y_i) - \sum_{i=1}^{k} \jmath_{\mathsf{S}}(s_i, d) \tag{342}$$

Weakening (330) using (331) and Lemma 1, we have

$$\epsilon' \geq \mathbb{E}\left[\min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[r(x^n, Y^n, S^n) \leq -\gamma \mid S^k\right] \cdot 1\left\{S^k \in \mathcal{T}_{k,n}\right\}\right] - \frac{1}{\sqrt{n+1}} \tag{343}$$

$$\geq \mathbb{E}\left[\mathbb{P}\left[r(x^{n\star}, Y^n, S^n) \leq -\gamma \mid S^k\right] \cdot 1\left\{S^k \in \mathcal{T}_{k,n}\right\}\right] - \frac{K}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \tag{344}$$

$$= \mathbb{P}\left[r(x^{n\star}, Y^n, S^n) \leq -\gamma,\ S^k \in \mathcal{T}_{k,n}\right] - \frac{K}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \tag{345}$$

$$\geq \mathbb{P}\left[r(x^{n\star}, Y^n, S^n) \leq -\gamma\right] - \mathbb{P}\left[S^k \notin \mathcal{T}_{k,n}\right] - \frac{K}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \tag{346}$$

$$\geq \mathbb{P}\left[r(x^{n\star}, Y^n, S^n) \leq -\gamma\right] - \frac{4C^2}{R^2(d)\bar{\Delta}^2}\frac{\mathcal{V}(d)}{k} - \frac{K}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \tag{347}$$

$$\geq \epsilon \tag{348}$$

where (344) is by Lemma 1, and (346) is by the union bound. To justify (348), observe that the quantities in Theorem 24 corresponding to the sum of independent random variables in (347) are

$$D_{n+k} = \frac{n}{n+k} I(\Pi(P_{\mathsf{X}^\star})) - \frac{k}{n+k} R(d) \tag{349}$$

$$\leq \frac{n}{n+k} C - \frac{k}{n+k} R(d) \tag{350}$$

$$V_{n+k} = \frac{n}{n+k} V(\Pi(P_{\mathsf{X}^\star})) + \frac{k}{n+k} \mathcal{V}(d) \tag{351}$$

$$\geq \frac{n}{n+k} V + \frac{k}{n+k} \mathcal{V}(d) - \frac{L|\mathcal{A}|^{\frac{1}{2}}}{2(n+k)} \tag{352}$$

$$T_{n+k} = \frac{n}{n+k} T(\Pi(P_{\mathsf{X}^\star})) + \frac{k}{n+k} \mathbb{E}\left[|\jmath_{\mathsf{S}}(\mathsf{S}, d) - R(d)|^3\right] \tag{353}$$

where the functions $\Pi(\cdot)$, $I(\cdot)$, $V(\cdot)$, $T(\cdot)$ are defined in (206), (248)–(250) in Appendix B. To show (352), recall that $V(P_{\mathsf{X}^\star}) = V$ by Property 4 in Appendix B and use (299) and Lipschitz continuity of $V(P_{\mathsf{X}})$ (Property 3 in Appendix B). Further, T_{n+k} is bounded uniformly in P_{X}, so (204) is upper bounded by some constant $B > 0$. Finally, using (350) and (352) in (368), we conclude that

$$-\gamma \geq (n+k) D_{n+k} - \sqrt{(n+k) V_{n+k}} Q^{-1}\left(\epsilon_{k,n}\right) \tag{354}$$

which enables us to lower bound the probability in (347) invoking the Berry-Esseen bound (Theorem 24). In view of (340), the resulting bound is equal to ϵ, and the proof of (348) is complete.

B. $V = 0$.

Fix an $0 < \alpha < \frac{1}{6}$.

If $\mathcal{V}(d) > 0$, we choose γ as in (331), and

$$\epsilon_{k,n} = \epsilon + \frac{B}{\sqrt{k}} + (n+1)^{|\mathcal{A}|-1} \exp\left(-\gamma\right) + \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} \tag{355}$$

where $B > 0$ is the same as in (340), and k, n are chosen so that both (328) and the following version of (329) hold:

$$nC - kR(d) \leq \sqrt{k\mathcal{V}(d)} Q^{-1}\left(\epsilon_{k,n}\right) - \gamma - \bar{\Delta} n^{\frac{1}{2}-\alpha} \tag{356}$$

where $\bar{\Delta} > 0$ was defined in Lemma 1. Weakening (330) using (212), we have

$$\epsilon' \geq \mathbb{P}\left[nC \leq \sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - \gamma - \bar{\Delta} n^{\frac{1}{2}-\alpha}\right] - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} - (n+1)^{|\mathcal{A}|-1} \exp\left(-\gamma\right) \tag{357}$$

$$\geq \mathbb{P}\left[\sqrt{k\mathcal{V}(d)} Q^{-1}\left(\epsilon_{k,n}\right) \leq \sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - kR(d)\right] - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} - (n+1)^{|\mathcal{A}|-1} \exp\left(-\gamma\right) \tag{358}$$

$$\geq \epsilon \tag{359}$$

where (358) uses (356), and (359) is by the Berry-Esseen bound.

If $\mathcal{V}(d) = 0$, which implies $\jmath_S(S_i, d) = R(d)$ a.s., we let

$$\gamma = (|\mathcal{A}| - 1)\log(n+1) - \log\left(1 - \epsilon - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}}\right) \tag{360}$$

and choose k, n that satisfy (328) and

$$kR(d) - nC \geq \gamma + \bar{\Delta} n^{\frac{1}{2}-\alpha} \tag{361}$$

Then, plugging $\jmath_S(S_i, d) = R(d)$ a.s. in (330), we have

$$\epsilon' \geq \min_{x^n \in \mathcal{A}^n} \mathbb{P}\left[\sum_{i=1}^{n} \imath_{\mathsf{X};\mathsf{Y}}(x_i; Y_i) \leq kR(d) - \gamma\right] - (n+1)^{|\mathcal{A}|-1} \exp\left(-\gamma\right) \tag{362}$$

$$\geq \mathbb{P}\left[nC \leq kR(d) - \gamma - \bar{\Delta} n^{\frac{1}{2}-\alpha}\right] - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} - (n+1)^{|\mathcal{A}|-1} \exp\left(-\gamma\right) \tag{363}$$

$$= 1 - \frac{1}{n^{\frac{1}{4}-\frac{3}{2}\alpha}} - (n+1)^{|\mathcal{A}|-1} \exp\left(-\gamma\right) \tag{364}$$

$$= \epsilon \tag{365}$$

where (363) invokes (212), (364) is by the choice of k, n, and (365) follows from the choice of γ.

C. Symmetric channel.

We show that if the channel is such that the distribution of $\imath_{\mathsf{X},\mathsf{Y}^\star}(\mathsf{x};\mathsf{Y}^\star)$ does not depend on the choice $\mathsf{x} \in \mathcal{A}$, Theorem 2 leads to a tighter third-order term than (111).

If either $V > 0$ or $\mathcal{V}(d) > 0$, let

$$\gamma = \frac{1}{2}\log n \tag{366}$$

$$\epsilon_{k,n} = \epsilon + \frac{B}{\sqrt{n+k}} + \frac{1}{\sqrt{n}} \tag{367}$$

where $B > 0$ can be chosen as in (340), and let k, n be such that the following version of (329) (with the remainder $\theta(n)$ satisfying (111) with $\underline{c} = \frac{1}{2}$) holds:

$$nC - kR(d) \le \sqrt{nV + k\mathcal{V}(d)}Q^{-1}\left(\epsilon_{k,n}\right) - \gamma \tag{368}$$

Theorem 2 and Theorem 24 imply that any (k, n, d, ϵ') code must satisfy, for an arbitrary sequence $x^n \in \mathcal{A}^n$,

$$\epsilon' \ge \mathbb{P}\left[\sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i, d) - \sum_{j=1}^{n} \imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) \ge \gamma\right] - \exp\left(-\gamma\right) \tag{369}$$

$$\ge \epsilon \tag{370}$$

If both $V = 0$ and $\mathcal{V}(d) = 0$, choose k, n satisfy

$$kR(d) - nC \ge \gamma \tag{371}$$

$$= \log\frac{1}{1-\epsilon} \tag{372}$$

Substituting (372) and $\jmath_{\mathsf{S}}(S_i, d) = R(d)$, $\imath_{\mathsf{X};\mathsf{Y}^\star}(x_i; Y_i) = C$ a.s. in (369), we conclude that the right side of (369) equals ϵ, so $\epsilon' \ge \epsilon$ whenever a (k, n, d, ϵ') code exists.

D. Gaussian channel

In view of Remark 9, it suffices to consider the equal power constraint (125). The spherically-symmetric $P_{\bar{Y}^n} = P_{Y^{n\star}} = P_{\mathsf{Y}^\star} \times \ldots \times P_{\mathsf{Y}^\star}$, where $\mathsf{Y}^\star \sim \mathcal{N}(0, \sigma_{\mathsf{N}}^2(1+P))$, satisfies the symmetry assumption of Theorem 2. In fact, for all $x^n \in \mathcal{F}(\alpha)$, $\imath_{X^n;Y^{n\star}}(x^n; Y^n)$ has the same distribution under $P_{Y^n|X^n=x^n}$ as (cf. (158))

$$G_n = \frac{n}{2}\log\left(1+P\right) - \frac{\log e}{2}\left(\frac{P}{1+P}\sum_{i=1}^{n}\left(W_i - \frac{1}{\sqrt{P}}\right)^2 - n\right) \tag{373}$$

where $W_i \sim \mathcal{N}\left(\frac{1}{\sqrt{P}}, 1\right)$, independent of each other. Since G_n is a sum of i.i.d. random variables, the mean of $\frac{G_n}{n}$ is equal to $C = \frac{1}{2}\log(1+P)$ and its variance is equal to (110), the result follows analogously to (366)–(370).

APPENDIX D
PROOF OF THE ACHIEVABILITY PART OF THEOREM 9

A. Almost lossless coding ($d = 0$) over a DMC.

The proof consists of an asymptotic analysis of the bound in Theorem 8 by means of Theorem 24. Weakening (105) by fixing $P_{X^n} = P^\star_{X^n} = P_{\mathsf{X}^\star} \times \ldots \times P_{\mathsf{X}^\star}$, we conclude that there exist a $(k, n, 0, \epsilon')$ code with

$$\epsilon' \leq \mathbb{E}\left[\exp\left(-\left|\sum_{i=1}^n \imath^\star_{\mathsf{X};\mathsf{Y}}\left(X_i^\star; Y_i^\star\right) - \sum_{i=1}^k \imath_{\mathsf{S}}(S_i)\right|^+\right)\right] \tag{374}$$

where $(S^k, X^{n\star}, Y^{n\star})$ are distributed according to $P_{S^k} P_{X^{n\star}} P_{Y^n|X^n}$. The case of equiprobable S has been tackled in [14]. Here we assume that $\imath_{\mathsf{S}}(\mathsf{S})$ is not a constant, that is, $\mathrm{Var}\left[\imath_{\mathsf{S}}(\mathsf{S})\right] > 0$.

Let k and n be such that

$$nC - kH(\mathsf{S}) \geq \sqrt{nV + k\mathcal{V}}\, Q^{-1}\left(\epsilon - \frac{B+1}{\sqrt{n+k}}\right) + \frac{1}{2}\log(n+k) \tag{375}$$

where $\mathcal{V} = \mathrm{Var}\left[\imath_{\mathsf{S}}(\mathsf{S})\right]$, and B is the Berry-Esseen ratio (204) for the sum of $n+k$ independent random variables appearing in the right side of (374). Note that B is finite due to:

- $\mathrm{Var}\left[\imath_{\mathsf{S}}(\mathsf{S})\right] > 0$;
- the third absolute moment of $\imath_{\mathsf{S}}(\mathsf{S})$ is finite;
- the third absolute moment of $\imath^\star_{\mathsf{X};\mathsf{Y}}(\mathsf{X}^\star; \mathsf{Y}^\star)$ is finite, as observed in Appendix B.

Therefore, (375) can be written as of (106) with the remainder therein satisfying (117). So, it suffices to prove that if k, n satisfy (375), the right side of (374) is upper bounded by ϵ. Let

$$\mathcal{T}_{k,n} = \Bigg\{ \left(s^k, x^n, y^n\right) \in \mathcal{S}^k \times \mathcal{A}^n \times \mathcal{B}^n : \\ \sum_{i=1}^n \imath^\star_{\mathsf{X};\mathsf{Y}}\left(x_i; y_i\right) - \sum_{i=1}^k \imath_{\mathsf{S}}(s_i) \geq nC - kH(\mathsf{S}) - \sqrt{nV + k\mathcal{V}}\, Q^{-1}\left(\epsilon - \frac{B+1}{\sqrt{n+k}}\right) \Bigg\} \tag{376}$$

By the Berry-Esseen bound (Theorem 24),

$$\mathbb{P}\left[\left(S^k, X^{n\star}, Y^{n\star}\right) \notin \mathcal{T}_{k,n}\right] \leq \epsilon - \frac{1}{\sqrt{n+k}} \tag{377}$$

We now further upper bound (374) as

$$\epsilon' \le \mathbb{E}\left[\exp\left(-\left|\sum_{i=1}^n \imath^\star_{\mathsf{X};\mathsf{Y}}\left(X_i^\star;Y_i^\star\right) - \sum_{i=1}^k \imath_{\mathsf{S}}(S_i)\right|^+ 1_{\mathcal{T}_{k,n}}\left(S^k, X^{n\star}, Y^{n\star}\right)\right)\right]$$

$$+ \mathbb{P}\left[\left(S^k, X^{n\star}, Y^{n\star}\right) \notin \mathcal{T}_{k,n}\right] \tag{378}$$

$$\le \exp\left(-\left|nC - kH(\mathsf{S}) - \sqrt{nV + k\mathcal{V}}Q^{-1}\left(\epsilon - \frac{B+1}{\sqrt{n+k}}\right)\right|^+\right)$$

$$+ \epsilon - \frac{1}{\sqrt{n+k}} \tag{379}$$

$$\le \epsilon \tag{380}$$

where (380) follows from (375).

B. Lossy coding over a DMC.

The proof consists of the asymptotic analysis of the bound in Theorem 7 using Theorem 24 and Lemma 5 below, which deals with asymptotic behavior of distortion d-balls. Note that Lemma 5 is the only step that requires finiteness of the ninth absolute moment of $\mathsf{d}(\mathsf{S}, \mathsf{Z}^\star)$ as required by restriction (iv).

Lemma 5 ([12, Lemma 2]). *Under restrictions* (ii)–(iv)*, there exist constants* $n_0, c, K > 0$ *such that for all* $n \ge n_0$*,*

$$\mathbb{P}\left[\log\frac{1}{P_{Z^{k\star}}(B_d(S^k))} \le \sum_{i=1}^k \jmath_{\mathsf{S}}(S_i, d) + \tilde{c}\log k + c\right] \ge 1 - \frac{K}{\sqrt{k}} \tag{381}$$

where $\tilde{c} = \bar{c} - \frac{3}{2}$*, with* $\bar{c}$ *given by* (114)*.*

We weaken (78) by fixing

$$P_{X^n} = P_{X^{n\star}} = P_{\mathsf{X}^\star} \times \ldots \times P_{\mathsf{X}^\star} \tag{382}$$

$$P_{Z^k} = P_{Z^{k\star}} = P_{\mathsf{Z}^\star} \times \ldots \times P_{\mathsf{Z}^\star} \tag{383}$$

$$\log M = kR(d) + 2k\Delta \tag{384}$$

$$\gamma = \frac{1}{2}\log_e k \tag{385}$$

where $\Delta > 0$, so there exists a (k, n, d, ϵ') code with

$$\epsilon' \leq \Bigg\{ \mathbb{E}\left[\exp\left\{-\left|\sum_{i=1}^n \imath^\star_{\mathsf{X};\mathsf{Y}}(X_i^\star; Y_i^\star) - \log \frac{\gamma H_M}{P_{Z^{k\star}}(B_d(S^k))}\right|^+\right\}\right] + \mathbb{E}\left[\left|e^{-\gamma} - \left(1 - P_{Z^{k\star}}(B_d(S^k))\right)^M\right|^+\right] + \mathbb{E}\left[\left(1 - P_{Z^{k\star}}(B_d(S^k))\right)^M\right] \Bigg\} \tag{386}$$

where $(S^k, X^{n\star}, Y^{n\star}, Z^{k\star})$ are distributed according to $P_{S^k} P_{X^{n\star}} P_{Y^n|X^n} P_{Z^{k\star}}$. We need to show that for k, n satisfying (106), the right side of (386) is bounded by ϵ. We begin by showing that the last two terms in (386) are negligible. Denote

$$U_k = \log M - \sum_{i=1}^k \jmath_{\mathsf{S}}(s_i, d) - \tilde{c} \log k - c \tag{387}$$

where $\tilde{c}$ and c are defined in Lemma 5. The second term in (390) is bounded as,

$$\mathbb{E}\left[\left|e^{-\gamma} - \left(1 - P_{Z^{k\star}}(B_d(S^k))\right)^M\right|^+\right] \leq e^{-\gamma} = \frac{1}{\sqrt{k}} \tag{388}$$

and the third term as,

$$\mathbb{E}\left[\left(1 - P_{Z^{k\star}}(B_d(S^k))\right)^M\right] \leq \mathbb{E}\left[e^{-MP_{Z^{k\star}}(B_d(S^k))}\right] \tag{389}$$

$$\leq \mathbb{E}\left[e^{-\exp(U_k)}\right] + \frac{K}{\sqrt{n}} \tag{390}$$

$$= \mathbb{E}\left[e^{-\exp(U_k)} 1\left\{U_k < \log \frac{\log_e k}{2}\right\}\right] + \mathbb{E}\left[e^{-\exp(U_k)} 1\left\{U_k \geq \log \frac{\log_e k}{2}\right\}\right] + \frac{K}{\sqrt{k}} \tag{391}$$

$$\leq \mathbb{P}\left[U_k < \log \frac{\log_e k}{2}\right] + \frac{1}{\sqrt{k}} \mathbb{P}\left[U_k \geq \log \frac{\log_e k}{2}\right] + \frac{K}{\sqrt{k}} \tag{392}$$

$$\leq \mathbb{P}\left[U_k < \log \frac{\log_e k}{2}\right] + \frac{K+1}{\sqrt{k}} \tag{393}$$

$$\leq \mathbb{P}\left[\sum_{i=1}^k \jmath_{\mathsf{S}}(S_i, d) \geq kR(d) + k\Delta\right] + \frac{K+1}{\sqrt{k}} \tag{394}$$

$$\leq \frac{1}{k} \frac{\mathcal{V}(d)}{(R(d) + \Delta)^2} + \frac{K+1}{\sqrt{k}} \tag{395}$$

where

- (390) is by Lemma 5;
- (392) upper bounds $e^{-\exp(U_k)}$ by 1 and $\frac{1}{\sqrt{k}}$, respectively;
- (394) holds for large enough k by the choice of M in (384);
- (395) is by Chebyshev's inequality.

Note that the reasoning leading up to (392) follows that in [12, (107)–(110)].

The first term in (386) is upper-bounded using Lemma 5 as:

$$\mathbb{E}\left[\exp\left\{-\left|\sum_{i=1}^{n} \imath^{\star}_{\mathsf{X};\mathsf{Y}}(X_i^{\star};Y_i^{\star}) - \log\frac{\gamma H_M}{P_{Z^{k\star}}(B_d(S^k))}\right|^{+}\right\}\right] \tag{396}$$

$$\leq \mathbb{E}\left[\exp\left\{-\left|U_{k,n}\right|^{+}\right\}\right] + \frac{K}{\sqrt{k}} \tag{397}$$

with

$$U_{k,n} = \sum_{i=1}^{n} \imath^{\star}_{\mathsf{X};\mathsf{Y}}(X_i^{\star};Y_i^{\star}) - \sum_{i=1}^{k} \jmath_{\mathsf{S}}(S_i,d) - \tilde{c}\log k - \log\left(\gamma H_M\right) - c \tag{398}$$

We first consider the (nontrivial) case when either $\mathcal{V}(d) + V > 0$. Let k and n be such that

$$nC - kR(d) \geq \sqrt{nV + k\mathcal{V}(d)}\,Q^{-1}\left(\epsilon_{k,n}\right) + \left(\tilde{c} + \frac{1}{2}\right)\log k + \log\gamma H_M + c \tag{399}$$

$$\epsilon_{k,n} = \epsilon - \frac{B}{\sqrt{n+k}} - \frac{1}{k}\frac{\mathcal{V}(d)}{(R(d)+\Delta)^2} + \frac{2K+3}{\sqrt{k}} \tag{400}$$

where constants c and $\tilde{c}$ are defined in Lemma 5, and B is the Berry-Esseen ratio (204) for the sum of $n+k$ independent random variables appearing in (397). Note that B is finite because:

- either $\mathcal{V}(d) > 0$ or $V > 0$ by the assumption;
- the third absolute moment of $\jmath_{\mathsf{S}}(\mathsf{S}, d)$ is finite by restriction (iv) as spelled out in (321);
- the third absolute moment of $\imath^{\star}_{\mathsf{X};\mathsf{Y}}(\mathsf{X}^{\star};\mathsf{Y}^{\star})$ is finite, as observed in Appendix B.

Since

$$H_M = \log_e M + O\left(1\right) \tag{401}$$

applying a Taylor series expansion to (399) with the choice of M and γ in (384) and (385), we conclude that (399) can be written as (106) with the remainder term satisfying (112).

It remains to further upper bound (397) using (399). Let

$$\mathcal{T}_{k,n} = \Bigg\{ \left(s^k, x^n, y^n\right) \in \mathcal{S}^k \times \mathcal{A}^n \times \mathcal{B}^n : \sum_{i=1}^{n} \imath^{\star}_{\mathsf{X};\mathsf{Y}}(x_i;y_i) - \sum_{i=1}^{k} \jmath_{\mathsf{S}}(s_i,d) \geq nC - kR(d) - \sqrt{nV + k\mathcal{V}(d)}\,Q^{-1}\left(\epsilon_{k,n}\right) \Bigg\} \tag{402}$$

By the Berry-Esseen bound (Theorem 24),

$$\mathbb{P}\left[\left(S^k, X^{n\star}, Y^{n\star}\right) \notin \mathcal{T}_{k,n}\right] \le \epsilon_{k,n} + \frac{B}{\sqrt{n+k}} \tag{403}$$

so the expectation in the right side of (397) is upper-bounded as

$$\begin{aligned} &\mathbb{E}\left[\exp\left\{-\left|U_{k,n}\right|^+\right\}\right] \\ &\le \mathbb{E}\left[\exp\left(-\left|U_{k,n}\right|^+ 1\left\{\left(S^k, X^{n\star}, Y^{n\star}\right) \in \mathcal{T}_{k,n}\right\}\right)\right] + \mathbb{P}\left[\left(S^k, X^{n\star}, Y^{n\star}\right) \notin \mathcal{T}_{k,n}\right] \qquad (404) \\ &\le \frac{1}{\sqrt{k}}\mathbb{P}\left[\left(S^k, X^{n\star}, Y^{n\star}\right) \in \mathcal{T}_{k,n}\right] + \epsilon_{k,n} + \frac{B}{\sqrt{n+k}} \qquad (405) \end{aligned}$$

where we used (399) and (402) to upper bound the exponent in the right side of (404).

Assembling (388), (395), (397) and (405), we conclude that the right side of (386) is indeed upper bounded by ϵ.

Finally, consider the case $V = \mathcal{V}(d) = 0$, which implies $\jmath_{\mathsf{S}}(\mathsf{S}, d) = R(d)$ and $\imath^\star_{\mathsf{X};\mathsf{Y}}(X_i^\star; Y_i^\star) = C$ almost surely, and let k and n be such that

$$nC - kR(d) \ge \tilde{c}\log k + \log\gamma H_M + c + \log\frac{1}{\epsilon - \frac{2K+2}{\sqrt{k}}} \tag{406}$$

where constants c and $\tilde{c}$ are defined in Lemma 5. Then

$$\mathbb{E}\left[\exp\left\{-\left|U_{k,n}\right|^+\right\}\right] \le \epsilon - \frac{2K+2}{\sqrt{k}} \tag{407}$$

which, together with (388), (395) and (397) implies that $\epsilon' \le \epsilon$, as desired.

C. *Lossy or almost lossless coding over a Gaussian channel*

In view of Remark 9, it suffices to consider the equal power constraint (125). As shown in the proof of Theorem 17, for any distribution of X^n on the power sphere,

$$\imath_{X^n;Y^n}(X^n; Y^n) \ge G_n - F \tag{408}$$

where G_n is defined in (373) (cf. (158)) and F is a (computable) constant.

Now, the proof for almost lossless coding can be modified to work for the Gaussian channel by adding $\log F$ to the right side of (375) and replacing $\sum_{i=1}^n \imath^\star_{\mathsf{X};\mathsf{Y}}\left(X_i^\star; Y_i^\star\right)$ in (374) and (378) with $G_n - \log F$, and in (376) with G_n.

The proof for lossy coding in Appendix D-B is adapted for the Gaussian channel by replacing all $\sum_{i=1}^n \imath^\star_{\mathsf{X};\mathsf{Y}}\left(X_i^\star; Y_i^\star\right)$ with G_n, and all H_M (except (401)) with $H_M F$.

APPENDIX E
PROOF OF THEOREM 21

Applying the Berry-Esseen bound to (179), we obtain

$$\begin{aligned} & D_1(n, \epsilon, \alpha) \\ \geq & \min_{\substack{P_{\mathsf{Z}|\mathsf{S}}: \\ I(\mathsf{S};\mathsf{Z}) \leq C(\alpha)}} \left\{ \mathbb{E}\left[\mathsf{d}(\mathsf{S},\mathsf{Z})\right] + \sqrt{\frac{\mathrm{Var}\left[\mathsf{d}(\mathsf{S},\mathsf{Z})\right]}{n}} Q^{-1}\left(\epsilon + \frac{B}{\sqrt{n}}\right) \right\} \\ & = D(C(\alpha)) + \sqrt{\frac{\mathcal{W}_1(\alpha)}{n}} Q^{-1}\left(\epsilon + \frac{B}{\sqrt{n}}\right) \end{aligned} \tag{409}$$

where B is the Berry-Esseen ratio, and (409) follows by the application of Lemma 3 with $\mathcal{D} = \left\{P_{\mathsf{SZ}} = P_{\mathsf{Z}|\mathsf{S}} P_{\mathsf{S}} \colon I(\mathsf{S};\mathsf{Z}) \leq C(\alpha)\right\}$, $\varphi_1 = 1$, $\varphi_2 = \frac{1}{\sqrt{n}}$. Note that $\mathbb{E}\left[d(\mathsf{S},\mathsf{Z})\right]$ is a linear function of P_{SZ} and $\sqrt{\mathrm{Var}\left[d(\mathsf{S},\mathsf{Z})\right]}$ is a continuously differentiable function of P_{SZ}, so conditions (223) and (225) hold with the metric being the usual Euclidean distance between vectors in $\mathbb{R}^{|\mathcal{S}|\times|\hat{\mathcal{S}}|}$. So, (409) follows immediately upon observing that by the definition of the rate-distortion function, $\mathbb{E}\left[\mathsf{d}(\mathsf{S},\mathsf{Z})\right] \geq \mathbb{E}\left[\mathsf{d}(\mathsf{S},\mathsf{Z}^\star)\right] = D(C(\alpha))$ for all $P_{\mathsf{Z}|\mathsf{S}}$ such that $I(\mathsf{S};\mathsf{Z}) \leq C(\alpha)$.

REFERENCES

[1] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE Int. Conv. Rec.*, vol. 7, pp. 142–163, Mar. 1959, reprinted with changes in *Information and Decision Processes*, R. E. Machol, Ed. New York: McGraw-Hill, 1960, pp. 93-126.

[2] I. Csiszár, “Joint source-channel error exponent,” *Prob. Contr. & Info. Theory*, vol. 9, no. 5, pp. 315–328, 1980.

[3] ——, “On the error exponent of source-channel transmission with a distortion threshold,” *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 823–828, 1982.

[4] R. Pilc, “Coding theorems for discrete source-channel pairs,” Ph.D. dissertation, M.I.T., 1967.

[5] ——, “The transmission distortion of a discrete source as a function of the encoding block length,” *Bell Syst. Tech. J.*, vol. 47, pp. 827–885, 1968.

[6] A. D. Wyner, “Communication of analog data from a Gaussian source over a noisy channel,” *Bell Syst. Tech. J.*, vol. 47, no. 5, pp. 801–812, May/June 1968.

[7] ——, “On the transmission of correlated Gaussian data over a noisy channel with finite encoding block length,” *Information and Control*, vol. 20, no. 3, pp. 193–215, April 1972.

[8] I. Csiszár, “An abstract source-channel transmission problem,” *Problems of Control and Information Theory*, vol. 12, no. 5, pp. 303–307, 1983.

[9] A. Tauste Campo, G. Vazquez-Vilar, A. Guillén i Fàbregas, and A. Martinez, “Random-coding joint source-channel bounds,” in *Proceedings 2011 IEEE International Symposium on Information Theory*, Saint-Petersburg, Russia, Aug. 2011.

[10] D. Wang, A. Ingber, and Y. Kochman, “The dispersion of joint source-channel coding,” in *49th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, Sep. 2011.

[11] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1147 – 1158, May 2003.

[12] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.

[13] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, no. 1, pp. 57–71, 1974.

[14] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[15] H. Poor, *An introduction to signal detection and estimation*. Springer, 1994.

[16] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 44 –54, Jan 1995.

[17] A. Dembo and I. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Transactions on Information Theory*, vol. 48, pp. 1590–1615, June 2002.

[18] D. Sakrison, "A geometric treatment of the source encoding of a Gaussian random variable," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 481–486, May 1968.

[19] F. Jelinek, *Probabilistic information theory: discrete and memoryless models*. McGraw-Hill, 1968.

[20] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge Univ Press, 2011.

[21] J. Goblick, T., "Theoretical limitations on the transmission of data from analog sources," *IEEE Transactions on Information Theory*, vol. 11, no. 4, pp. 558 – 567, Oct. 1965.

[22] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 136–152, Jan. 2000.

[23] W. Feller, *An introduction to probability theory and its applications*, 2nd ed. John Wiley & Sons, 1971, vol. II.